ELSEVIER

Chemical Engineering Science 61 (2006) 966–988

Chemical Engineering Science

www.elsevier.com/locate/ces

Advances in protein structure prediction and de novo protein design: A review

C.A. Floudas*, H.K. Fung, S.R. McAllister, M. Mönnigmann, R. Rajgaria

Department of Chemical Engineering, Princeton University, Princeton, NJ 08544-5263, USA

Received 18 November 2004; received in revised form 1 April 2005; accepted 2 April 2005
Available online 4 November 2005

Abstract

This review provides an exposition to the important problems of (i) structure prediction in protein folding and (ii) de novo protein design. The recent advances in protein folding are reviewed based on a classification of the approaches in comparative modeling, fold recognition, and first principles methods with and without database information. The advances towards the challenging problem of loop structure prediction and the first principles method, ASTRO-FOLD, along with the developments in the area of force-fields development have been discussed. Finally, the recent progress in the area of de novo protein design is presented with focus on template flexibility, in silico sequence selection, and successful peptide and protein designs.

Keywords: Protein folding; De novo protein design; Loop structure prediction; Force field development

1. Introduction

Proteins are linear chains of amino acids that adopt a unique three-dimensional structure in their native surroundings. It is this native structure that allows the protein to carry out its biochemical function. Levinthal's paradox (Levinthal, 1969; Zwanzig et al., 1992) raised the question why and how a sequence of amino acids can fold into its functional native structure given the abundance of geometrically possible structures.

The pioneering experiments of Anfinsen (1973) shed light on this problem. According to Anfinsen's thermodynamic hypothesis, proteins are not assembled into their native structures by a biological process, but folding is a purely physical process that depends only on the specific amino acid sequence of the protein and the surrounding solvent. Anfinsen's hypothesis implies that in principle protein structure can be predicted if a model of the free energy is available, and if the global minimum of this function can be identified. This idea defines the protein structure prediction problem well, as it allows to infer macroscopic structure of many proteins from a few types of microscopic interactions between the protein's constituents. On the other hand, protein structure prediction remains utterly complex, since even short amino acid sequences can form an abundant number of geometric structures among which the free energy minimum has to be identified.

A protein is composed of several levels of structure. The primary structure of a protein is described by the specific amino acid sequence. Additionally, patterns of local bonding can be identified as secondary structure. The two most common types of secondary structure are α-helices and β-sheets. Connecting these elements of secondary structure are loop regions. The tertiary structure is then the final three-dimensional structure of these elements after the protein folds into its native state. Fig. 1 illustrates an example protein structure.

The protein structure prediction problem is a fundamental problem treated across disciplines. From a chemical engineering point of view, the structure prediction problem is of great interest, because it is a prerequisite for successfully

* Corresponding author. Tel.: +1 609 258 4595; fax: +1 609 258 0211.
E-mail address: floudas@titan.princeton.edu (C.A. Floudas).

0009-2509/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ces.2005.04.009

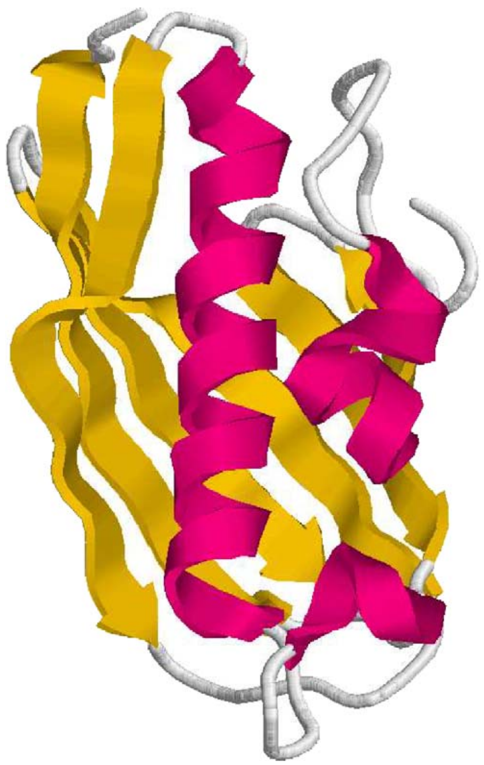

Fig. 1. The three-dimensional structure of a protein. α-helices are colored pink and the β-strands are shown as yellow arrows. (PDB code:1Q4S) All protein images are generated with the RASMOL molecular visualization package (Sayle and Milner-White, 1995).

tackling de novo protein design. In de novo protein design the ultimate objective is to identify amino acid sequences that fold into proteins with desired functions. De novo protein design can be looked upon as a product design problem on the molecular scale.

Many approached to computational protein structure prediction using first principles have been developed over the last decade that are based on Anfinsen's thermodynamic hypothesis. Section 2 attempts to give an overview of recent developments. Computational structure prediction based on first principles is, however, not the only way to determine protein structure. The number of protein structures that have been determined experimentally continues to grow rapidly. At the end of 2004, the number of structures freely available from the Protein Data Bank (Berman et al., 2000) is approaching 28,000. The availability of experimental data on protein structures has inspired the development of methods for computational structure prediction that are knowledge-based rather than physics based. In contrast to methods that attempt to minimize the free energy and derive the structure from first principles, these knowledge-based approaches search databases of known structures to infer information about an amino acid sequence of unknown three-dimensional structure. While such database methods have been criticized for not helping to obtain a fundamental understanding of the mechanisms that drive structure formation, these knowledge-based methods can often successfully predict unknown three dimensional structures.

Progress for all variants of computational protein structure prediction methods is assessed in the biannual, community-wide Critical Assessment of Protein Structure Prediction (CASP) experiments (Moult et al., 2003, 2001, 1997; Moult, 1999). In the CASP experiments, research groups are invited to apply their prediction methods to amino acid sequences for which the native structure is not known but to be determined and to be published soon. Even though the number of amino acid sequences provided by the CASP experiments is small, these competitions provide a good measure to benchmark methods and progress in the field in an arguably unbiased manner (Murzin, 2004). The overview on computational protein structure prediction methods given in this review will draw on results from recent CASP experiments.

Research on protein structure prediction methods as witnessed in the biannual CASP experiments has been motivated to a large extent by scientific curiosity. Protein structure prediction is, however, not only interesting from a *scientific*, but also from an *engineering* point of view. It constitutes a major part of the de novo protein design problem, which is also called the inverse protein folding problem (Pabo, 1983; Drexler, 1981) that requires the determination of an amino acid sequence compatible with a given three-dimensional structure. De novo protein design problem is the "inverse" of the protein folding problem because it starts with the structure rather than the sequence and looks for all sequences that will fold into such structure. Experimentalists have tackled this problem with mutagenesis, rational design, and directed evolution. These methods are, however, restricted with respect to the number of mutant structures that can be screened experimentally which is typically in the range of 10^3–10^6 sequences (Voigt et al., 2001). Computational protein design methods, in contrast, allow for the screening of overwhelmingly large parts of sequence space. Toward this end, the paper summarizes recent progress in the field of de novo protein design.

The review is organized as follows. Section 2 provides an overview on methods for protein structure prediction and summarizes recent developments in all categories of approaches to the problem. Section 3 is devoted to methods for loop structure prediction. Section 4 outlines the first principles protein structure prediction method, ASTRO-FOLD. Section 5 discusses advances in force field development as they pertain to fold recognition and de novo protein design. Section 6 focuses on recent progress in de novo protein design.

2. Protein structure prediction

Numerous different approaches to protein structure prediction exist. Methods for structure prediction can be divided into four groups: (1) comparative modeling; (2) fold recognition; (3) first principles methods with database information; and (4) first principles methods without database information. As prediction methods became more sophisticated, the boundaries between these categories have been blurred, and today methods exist that cannot clearly be appointed to any of the four categories. Despite the blurry transition between the categories a classification of methods is helpful, and the above categories provide a reasonable starting point. In this section we will give a brief overview of methods in the four categories and point out where the categories overlap. Different points of view on first principles methods will be taken into account.

2.1. Comparative modeling

In comparative modeling the structure of a protein is predicted by comparing its amino acid sequence to sequences for which the native three-dimensional structure is already known. Comparative modeling is based on the observation that sequence similarity implies structural similarity. The accuracy of predictions by comparative modeling, however, strongly depends on the degree of sequence similarity. If the target and the template share more than 50% of their sequences, predictions usually are of high quality and have been shown to be as accurate as low-resolution X-ray predictions (Kopp and Schwede, 2004). For 30–50% sequence identity more than 80% of the C^{α}-atoms can be expected to be within 3.5 Å of their true positions (Kopp and Schwede, 2004), while for less than 30% sequence identity, the prediction is likely to contain significant errors (Kopp and Schwede, 2004; Vitkup et al., 2001).

Assessors of homology methods in CASP5 point out that the general approach to structure prediction by homology has not changed over the last two decades (Tramontano and Morea, 2003). In general, homology modeling consists of the selection of one or more templates from a database, their alignment to the target sequence, and refinement of side-chain geometry and regions of low sequence identity. Well-established algorithms and implementations for the first step, template identification, exist. BLAST (Altschul et al., 1990) and refinements such as PSI-BLAST (Altschul et al., 1997) are established to a degree that they serve as benchmarks to any new approach.

In a recent summary of the progress in homology modeling (Tramontano and Morea, 2003) it was pointed out that for easy targets the difference between average and best predictions by homology methods is modest only. Since approaches to homology modeling hardly differ with respect to template selection and alignment, this indicates that for easy targets the refinement steps are less crucial. While for hard targets the refinement step seems to be more important, it has been noted that refined models are typically not better than the template by more than 0.5 Å (Tramontano and Morea, 2003).

While there seems to be little progress in refining templates, continually improving sequence comparison techniques have broadened the scope of homology modeling. While 30% sequence similarity was considered to be the threshold for successful comparative modeling, predictions for targets with as low as 17% sequence similarity were made during the CASP4 experiment. During CASP5 sequence identity was as low as 6% (Tramontano and Morea, 2003), though prediction quality generally deteriorates as the sequence similarity drops. Recent progress in sequence comparison can be ascribed to the use of hidden Markov models (Karplus et al., 1998, 1999) and multiple sequence alignment (Notredame, 2002).

Advocates of comparative modeling claim that the importance of these methods will continue to grow as the number of experimentally determined structures grows steadily and, therefore, the number of sequences that can be related to a known structure is growing. This idea gave rise to the field of structural genomics. The central goal of structural genomics is to focus experimental efforts on representative structures for all anticipated folds (see Vitkup et al., 2001, and references therein). Estimates of the number of representative structures differ, however. Optimistic guesses anticipate that within the next 5–10 years, comparative modeling will be applicable to most sequences (Fiser et al., 2002). Structural genomics efforts exist around the world today. For a recent overview the reader is referred to Liu et al. (2004).

2.2. Fold recognition

While similar sequence implies similar structure, the converse is in general not true. In contrast, similar structures are often found for proteins for which no sequence similarity to any known structure can be detected. Fold recognition methods are one class of methods that aim at predicting the three-dimensional folded structure for amino acid sequences for which comparative modeling methods provide no reliable prediction.

Fold recognition methods are motivated by the notion that structure is evolutionary more conserved than sequence. As a consequence, the repertoire of different folds is more limited than suggested by sequence diversity. The number of different folds has been estimated based on clustering the structures deposited in the protein data bank (Berman et al., 2000) into families. Estimates vary strongly (Chothia, 1992; Orengo et al., 1994; Liu and Rost, 2002). According to a recent assessment, the protein data bank already contains enough structures to cover small protein structures up to a length of about a hundred residues (Kihara and Skolnick, 2003). While this evaluation may be optimistic, the number of existing folds can be expected to be orders of magnitudes smaller than the number of different sequences, since the number of different folds is only 800 according to the SCOP database (http://scop.berkeley.edu, November 2004).

Since the number of sequences is much larger than the number of folds, fold recognition methods attempt to identify a model fold for a given target sequence among the known folds even if no sequence similarity can be detected. Techniques for fold recognition include advanced sequence comparison methods. The boundary between fold recognition and homology modeling methods is blurry for this class of methods, and methods based on hidden Markov models (Karplus et al., 1999) and position specific iterated BLAST searches (Altschul et al., 1997) that have been mentioned in the above paragraph on comparative modeling are often considered fold recognition methods (Kim et al., 2003).

Another approach to fold recognition is based on secondary structure prediction and comparison. This subclass of methods is based on the observation that secondary structure similarity can exceed 80% for sequences that exhibit less than 10% sequence similarity (Klepeis and Floudas, 2003c). Clearly any such approach can only be as good as the underlying secondary structure prediction method. Concomitant to improvements in quality of secondary structure prediction, the use of secondary structure information has become more popular over the recent years. While secondary structure prediction seemed to stall at 60% of correctly assigned helices, strands and loops in the 1990s (Rost, 2001), amino acids subsequences can nowadays be reliably assigned to the structural types helix, strand or loop in more than 75% of the cases (Jones, 1999b; Rost, 2001; Przybylski and Rost, 2004). Improvements can be accredited to replacing single representative sequences by family profiles generated by PSI-BLAST, and to being able to include evolutionary diverse proteins in these families due to the availability of larger databases. As in other areas of protein structure prediction, secondary structure prediction has benefited from consensus methods that combine several different approaches into a single, generally more reliable, result (An and Friesner, 2002; Cuff et al., 1998).

Secondary structure information is often combined with other one-dimensional descriptors in fold recognition methods (e.g., with simple scores for solvent accessibility of each amino acid). Przybylski and Rost (2004) showed that a method based on secondary structure and solvent accessibility can outperform methods that take information on three-dimensional structure into account. Przybylski and Rost (2004) further showed that their approach performed better if information on known folds (e.g., information on secondary structure from experimental structure determination) was ignored deliberately. Their approach is based on predicting one dimensional descriptors for a target, and identifying a similar fold by comparing these descriptors to the descriptors of known folds. Suprisingly, the authors found the best results when comparing the target descriptors to the *predicted* descriptors of known folds, as opposed to using the experimentally-determined predictors of the known folds. In this sense, the fold recognition approach developed by Przybylski and Rost (2004) performed better if information on known folds (e.g., information on secondary structure from experimental structure determination) was ignored.

Finally, threading, that is, testing the compatibility of sequences with a known three-dimensional fold, is an important representative of fold recognition methods. Threading methods attempt to fit a target sequence to a known structure in a library of folds. Threading-based methods are known to be computationally expensive. In fact, globally optimal protein threading is known to be NP-hard (Garey and Johnson, 1979; Lathrop, 1994) if all residue–residue contacts are to be taken into account. Several threading methods ignore pairwise interaction between residues (Shi et al., 2001; Kelley et al., 2000; Jones, 1999a). In doing so, the threading problem is simplified considerably, and the simplified problem can be solved with dynamic programming (Bertsekas, 1995; Jones et al., 1992). In early methods of this kind, a one-dimensional string of features was recorded for known folds and compared to the target sequence (Bowie et al., 1991). The recorded features comprise attributes like buried side-chain area, side chain area covered by polar atoms including water, and the local secondary structure. In this manner, the three-dimensional structure of known proteins is converted into a one-dimensional sequence of descriptors. Models for the target structure are identified by seeking the most favorable alignment of the one-dimensional sequence of descriptors to any of those descriptor strings by dynamic programming (Bowie et al., 1991). Similar approaches use pair interaction potentials that describe a mean force derived from a database of known structures (Jones, 1999a; Xu and Xu, 2000; Kim et al., 2003; Jones et al., 1992). (See also Flöckner et al., 1997 and references therein.)

Skolnick and coworkers developed and successfully applied threading methods in the CASP experiments (Skolnick et al., 2003). Their most recent advance (Skolnick et al., 2004) is an iterative approach that first aligns target and known structures ignoring pairwise residue interactions. In subsequent alignments information from previous alignments is then used to evaluate pairwise interaction energies. The method combines three different pair potentials to account for the fact that different scoring functions are capable of assigning different target sequences to the same template. By identifying structurally similar regions in multiple templates, accurate regions of structure prediction can be distinguished from less accurate ones. Skolnick et al. (2004) found that accurate fragments can be identified even if no template is convincing as a whole. This observation led to the development of a fragment assembly method based on their threading approach (Zhang and Skolnick, 2004a,b). This fragment assembly approach is further discussed in the subsection on first principles prediction methods with database information.

Xu and Xu (2000) developed a threading algorithm that considers pair contacts between α-helices and β-strands and allows for alignment gaps in loop regions. They used a partitioning of template structures into helices and strands with open links that represent possible contacts to other regions. The approach finds a globally optimal recombination of regions and contacts if an upper bound on the spatial distance between contacting residues can be assumed. The method furthermore allows the incorporation of constraints about a target protein (e.g., known disulfide bonds or distance restraints) (Xu and Xu, 2000). More recently, they suggested to run a threading stage without pairwise potentials first. From this first stage, residue–residue contacts are inferred, and pairwise energy is taken into account in the second stage (Kim et al., 2003). An alternate approach poses the fold recognition problem as a global optimization of an energy function (Xu et al., 2003). This method accounts for the conservation of secondary structure through the definition of secondary structure cores, while still allowing for insertions or deletions. The energy function is defined as a weighted sum of pairwise interactions, mutations, gap penalties and secondary structure conservation. Although originally formulated as an integer programming problem, it is claimed that the linear programming relaxation frequently provides integer solutions. This method performed well in large-scale benchmarking tests (Xu and Li, 2003).

2.3. First principles prediction with database information

In the CASP experiments, the term ab initio has been used in a broader sense. The term refers to methods for structure prediction that do not use experimentally known structures. This use of the term ab initio has become more vague ever since the introduction of fragment based methods. These methods do not compare a target to a known protein, but they compare fragments, that is, short amino acid subsequences, of a target to fragments of known structures obtained from the Protein Data Bank (Berman et al., 2000). Once appropriate fragments have been identified, they are assembled to a structure, often with the aid of scoring functions and optimization algorithms. Since scoring functions resemble energy functions, and since fragment assembly with optimization algorithms resembles free energy optimization, this type of method bears some analogy to physics-based first principle methods. Clearly, however, fragment assembly methods cannot be considered first principle structure prediction methods in the same strict sense as first principle methods that are based on free energy minimization.

Fragment assembly methods are based on the premise that local interactions create a bias but do not uniquely define local structure. Local degrees of freedom are assumed to be fixed by non-local interactions, such as sheet formation or side chain interactions between non-neighboring residues, to result in a compact overall conformation. Fragment based methods approximate this structural bias by averaging over observed fragment geometries in known protein structures. Once appropriate fragments have been identified, compact structures can be assembled by randomly combining fragments using simulated annealing (Simons et al., 1997; Rohl et al., 2004). Subsequently, the fitness of a conformation can be assessed with scoring functions derived from conformational statistics of known proteins. Incorporation of information from independently conducted secondary structure predictions has resulted in improved scoring functions (Simons et al., 1999). Much of the progress in fragment assembly methods can be accredited to Baker and coworkers (see Rohl et al., 2004 and references therein). These fragment assembly methods performed consistently well across target classes in the recent CASP experiments (Aloy et al., 2003; Bradley et al., 2003).

Several other fragment-based methods performed well in the CASP experiments. Karplus and coworkers used a genetic algorithm to assemble fragments of nine residues with a cost function that maximizes residue burial (Karplus et al., 2003). Unlike fragment-based methods that evolved from threading, this method does not fix a single alignment or core residues, but the genetic algorithm even permits breaks in the backbone. Jones and coworkers (Jones and Guffin, 2003; Jones, 1997, 2001) predicted secondary structure for a target sequence with standard methods (Jones, 1999b) and selected fragments from a library of known folds that have the same two or three secondary structure elements as any corresponding sequence of secondary structural elements in the target. Fragments were reassembled using a scoring function with terms for long range interactions, short range interactions, solvation, steric clashes and hydrogen bonds by simulated annealing. Shao and Bystroff (2003) combined hidden Markov models for the detection of local sequence structure correlations and the derivation of contact potentials from contact maps for templates that aligned with a given target. A contact map is a symmetrical two-dimensional projection of the intraprotein distances that allows for easy identification of secondary structure elements. Skolnick, Kolinski and coworkers (Skolnick et al., 2001,2003, Zhang and Skolnick, 2004a,b) developed approaches that combine multiple sequence comparison, threading, optimization with scoring functions, and clustering. The method uses a united atom lattice model, which represents each residue by three or fewer atoms on a lattice (Zhang et al., 2003). Threading is used to provide information on long range interactions by identifying contacts between distant side-chains. They pointed out that the number of correctly predicted folds increases when threading information is incorporated, and that proteins of more than 120 residues in length can practically never be predicted correctly by their ab initio method without the use of information on long-range interactions obtained from threading. Furthermore, they denoted that clustering and selecting centroids of most populated clusters results in conformers closer to native than the lowest energy conformers (Zhang and Skolnick, 2004c). Lee and coworkers (Lee et al., 2004) compared the predicted secondary structure for fragments of 15 residues centered at each residue of the target structure with a fragment library. For each of the overlapping subsequences of length 15, the 20 most similar fragments from the library are recorded. Random conformations for the target structure are built up by looping over the residue sequence from the N- to the C-terminal, choosing fragments for each residue from the recorded ones, and identifying fragments with similar dihedral angles in their overlapping regions. Conformations are optimized by a stochastic genetic algorithm, conformational space annealing (Lee et al., 1997), using a scoring function that accounts for steric clashes and hydrogen bonds. They pointed out that the scoring function is crude (Lee et al., 2004). Nevertheless the performance of this method is considered promising in the most recent CASP meeting (Lee et al., 2004).

Several methods have focused on the assembly of knowledge-based secondary structure prediction into a native-like tertiary structure. A hierarchical approach to structure prediction has been proposed that iterates using a lattice model that becomes increasingly detailed with each step (Xia et al., 2000). Another class of methods introduces information on secondary structure and selected tertiary restraints and uses the principles of the deterministic αBB global optimization method (Adjiman et al., 1996, 1997, 1998a,b; Androulakis et al., 1995; Floudas, 2000) in combination with a reduced force field model (Eyrich et al., 1999a,b).

2.4. *First principles without database information*

Methods of this type make direct use of Anfinsen's thermodynamic hypothesis in that they attempt to identify the minimum of the free energy of the protein in its environment. Even though these methods are computationally demanding, first principle structure prediction is an indispensable complementary approach to any knowledge-based approach for several reasons. First, in some cases, even a remotely related structural homologue may not be available. In these cases, first principle methods are the only alternative. Second, new structures continue to be discovered which could not have been identified by methods which rely on comparison to known structures. Third, knowledge-based methods have been criticized for predicting protein structures without having to obtain a fundamental understanding of the mechanisms and driving forces of structure formation. First principle structure prediction methods, in contrast, base their predictions on physical models for these mechanisms. As such, they can therefore help to discriminate correct from incorrect modeling assumptions, and to deepen the understanding of the mechanisms of protein folding.

This class of methods can be applied to any given target sequence using only physically meaningful potentials and atom representations. With such a broad range of targets and the inability to directly or indirectly apply database information, these methods are the most difficult of the protein structure prediction methods. One hierarchical approach to structure prediction (LINUS) emphasizes the important role of local steric effects and conformational entropy (Srinivasan and Rose, 1995, 2002). Using a Metropolis Monte Carlo algorithm, the approach identifies protein conformational biases through a discrete set of moves and a simplified physics-based force field.

Scheraga and co-workers introduced a pioneering hierarchical approach to this problem that uses a simplified united-residue force field for initial calculations and then subsequent refinement of the coarse model using an all-atom potential (Pillardy et al., 2001; Lee et al., 2001; Liow et al., 2002, 1997a,b, 2001). This united-residue force field, which reduces the representation of each amino acid residue to just two interaction sites, allows the stochastic conformational space annealing algorithm to more efficiently identify low energy structures (Lee et al., 1998, 1997, 2000; Lee and Scheraga, 1999; Ripoll et al., 1998). Recent work has focused on improved algorithms to handle β-strands (Czaplewski et al., 2004a) and detailed analysis of the role of disulfide bonds in protein structure (Czaplewski et al., 2004b).

Floudas and co-workers introduced a first principles physics-based method, ASTRO-FOLD (Klepeis and Floudas, 2003c), which combines the classical and new views of protein folding. This approach begins by identifying helical regions through detailed free energy calculations and the application of global optimization methodologies (Klepeis and Floudas, 2002). The β-strands and β-sheet topologies can then be solved using a mixed-integer linear optimization formulation to maximize hydrophobic interactions (Klepeis and Floudas, 2003a). After using the secondary structure predictions to develop restraints, the tertiary structure is identified using a novel class of hybrid global optimization algorithms (Klepeis et al., 2003a,b). The details of the ASTRO-FOLD framework will be described further in Section 4. A recent validation of the ASTRO-FOLD approach on a double blind study of a 102 amino-acid protein is discussed in Klepeis et al. (2005).

3. Loop structure prediction

Ab initio methods have recently received increased attention in the prediction of loops, that is, those structures that join β-strands and helices in proteins. Loops exhibit greater structural variability than strands and helices, since they are often exposed at the surface of a protein and have relatively few contacts with the remainder of the structure. Loop structure therefore is considerably more difficult to predict than the structure of the geometrically highly regular strands and helices. From the observation that loops typically are no longer than 12 residues (Fiser et al., 2000) one may infer that loops are of lesser importance than the remaining parts of a protein. However, without loops and their structural flexibility, a protein cannot fold into a compact structure, and loops are often exposed to the surface of proteins and contribute to active and binding sites (Fiser et al., 2000). Consequently, loop structure and its contribution to protein function is of major importance.

In fold recognition methods, loops are often one of the limiting factors toward higher resolution of predicted structures (Jacobson et al., 2004; Li et al., 2004). The need for higher precision predictions of loops in comparative modeling has motivated recent research on first principles-based methods for loop structure prediction (Jacobson et al., 2004; Li et al., 2004; DePristo et al., 2003; Rohl et al., 2004).

In this section, we highlight recent developments and progress in first principles-based methods for loops. A brief but comprehensive summary of the methods that have been developed for loops prior to 2000 can be found in Fiser et al. (2000).

Baker and coworkers (Rohl et al., 2004) successfully used their fragment buildup method for loop prediction in CASP4. By building conformations from smaller fragments derived from databases, the problem of inadequate sampling encountered in other database methods is claimed to be overcome. Candidate loop structures are minimized by simulated annealing with respect to a heuristic scoring function that combines sequence similarity, secondary structure similarity for the residues adjacent to the loop, and geometric fit into the overall protein structure which is assumed to be known. The authors found that their approach yields similar but slightly worse results than a method that is based on physical energy function minimization and a consensus hybrid approach (Rohl et al., 2004). Deane and Blundell (2001) developed a consensus method that combines two sources of polypeptide fragments. Both a database of loops from known structures and a database of representative computer-generated fragments up to 12 residues long are used to identify loop structures that fit into a given protein structure. The quality of a candidate loop structure is evaluated based on geometric fit of the loop into the anchor regions in the protein. Zhang et al. (2003) developed a statistical potential function that compares favorably to recent approaches (de Bakker et al., 2003; DePristo et al., 2003) that use physical energy functions. Specifically, the authors find slightly worse predictions for loops of lengths up to 8 residues, but their predictions of loops of 9–12 residues is slightly better than recent predictions based on physical energy functions. Fiser et al. (2000) used a combination of a physical energy function and a scoring function that takes statistical preferences for dihedral angles and non-bonded atomic contacts into account. The resulting energy function is minimized with a combination of local optimization, molecular dynamics and simulated annealing. These authors improved the precision of loop predictions significantly compared to previous works, and they anticipated that further improvements would hinge upon the quality of energy and scoring functions rather than the ability to sample the conformational space of loops.

Recent progress in loop structure prediction has been achieved with approaches that combine dihedral angle sampling, steric clash detection, clustering, and scoring or energy function evaluation to build up ensembles of loop conformations, and to select representative structures from those ensembles. de Bakker and coworkers (de Bakker et al., 2003; DePristo et al., 2003) generated on the order of 10^3 candidate loop geometries by sampling protein backbone angle distributions that have been constructed from loops in known structures. During the sampling, loop conformations that result in steric clashes are filtered out. Similarly, candidate loops that do not fit into the overall protein geometry are discarded. Clustering approaches such as K-means (Hartigan and Wong, 1979) are used to classify candidate loops into groups of structures with similar geometry. By selecting representatives for groups, the number of loop candidates can be decreased, and redundancy in the ensembles can be reduced. The authors investigated different scoring and energy functions in detail (de Bakker et al., 2003) and found that a molecular mechanics force field outperformed a statistical potential in identifying low root mean square deviation (rmsd) structures from the ensemble.

Jacobson et al. (2004) developed a dihedral angle sampling approach similar to that of de Bakker and coworkers. Sampled loop structures are discarded if they contain steric clashes, if insufficient space exists for side chains, if loops travel too far away from the remainder of the protein, or if loops ends do not fit into the remainder of the protein. Before side chain optimization is applied, a K-means clustering approach (Hartigan and Wong, 1979) is used to pick representants from ensembles that comprise up to 10^6 conformers. The rmsds reported by this group of about 0.5, 1.0, and 2.5 Å for 5, 8, and 11 residue loops, respectively, are the lowest ones to date.

Both the de Bakker et al. (2003) and the Friesner group (Jacobson et al., 2004) note that the sampling process generates structures of considerably lower rmsd to the native structure than the lowest energy structures. Similarly, structures with lower energies than the native structure are found in the sampling process (Jacobson et al., 2004). While the use of physical energy functions is superior to statistical potential functions, these findings indicate that the accuracy of physical energy functions can be improved further (Jacobson et al., 2004).

Forrest and Woolf (2003) applied loop structure prediction to membrane protein loops. They combined Monte Carlo sampling and multi-temperature molecular dynamics to generate sets of conformations that are close to, but different from the native conformation. By applying several different energy functions to these test sets, they evaluated which energy contributions dominate the folding of membrane loops and which contributions can be neglected. They concluded that a complex description of the membrane itself is not necessary to predict membrane protein loop structure, and that it is crucial to account for solvation energy (Forrest and Woolf, 2003).

Zhang et al. (2003) developed a new statistical potential that compares favorably to physics-based potentials. They claimed that their statistics-based potential does not suffer from common limitations of scoring functions. In particular, their potential does not need to distinguish between buried and exposed residues a priori, but it can quantitatively predict the likelihood of a residue to be buried (Zhang et al., 2003). They tested their potential on the ensembles used previously by Forrest and Woolf (2003), DePristo et al. (2003)

and Jacobson et al. (2004). The results suggest that the statistical energy function can provide predictions comparable in accuracy to physical energy functions for two to eight residues, and better predictions for nine to twelve residues.

Mönnigmann and Floudas (2005) investigated the loop structure prediction problem with flexible stems. This problem is considerably more difficult than the loop reconstruction problem treated before (Jacobson et al., 2004; de Bakker et al., 2003; Xiang et al., 2002) in that neither the loop anchor geometry nor the overall protein geometry are assumed to be known. Mönnigmann and Floudas (2005) used a dihedral angle sampling approach similar to those used by Jacobson et al. (2004) and de Bakker et al. (2003) to build up ensembles of 2000 conformers. Loops structures were optimized with a first principles energy function, and a new iterative clustering approach was applied to filter out conformers that are far from the native structure. As opposed to previous approaches, clustering was not used to group conformers and identify representative conformers that are close to native, but to identify conformers that are far from native. By discarding these conformers, the quality of the ensembles could be improved. Mönnigmann and Floudas (2005) compared different methods of selecting conformers from the ensembles. They found that cluster size after iterative clustering outperforms energy and colony energy (Xiang et al., 2002). They applied their methodology to more than 3300 loops ranging from 10 residues (4 loop residues and 3 stem residues at both ends) to 20 residues (14 loop residues and 3 stem residues at both ends). Rmsds ranged from 2.65 Å for 10 residues to 5.04 Å for 20 residues with an approximately linear dependence of rmsd on loop length.

4. ASTRO-FOLD protein structure prediction approach

One successful prediction method is the first principles ASTRO-FOLD protein folding approach developed by Floudas and coworkers (Klepeis and Floudas, 2003c). The main thrusts of this approach are (1) α-helical prediction through detailed free energy calculations (Klepeis and Floudas, 2002), (2) a mixed-integer linear optimization formulation for the β-sheet prediction (Klepeis and Floudas, 2003a; Floudas, 1995), (3) derivation of secondary structure restraints and loop modeling, and (4) the application of the αBB global optimization algorithm (Adjiman et al., 1996, 1997, 1998a,b; Androulakis et al., 1995; Floudas, 2000) to tertiary structure prediction.

4.1. α-helix prediction

The first stage of the ASTRO-FOLD method applies the principles of hierarchical folding to the prediction of α-helical regions in proteins (Klepeis and Floudas, 2002). The application of hierarchical folding to α-helix prediction is justified by observations of the rapid formation of native-like helical segments. One proposed mechanism of the helix–coil transition suggests helical nucleation and propagation is based on local interactions (Honig and Yang, 1995).

This first principles prediction of α-helical segments of proteins begins by dividing the amino acid sequence into a series of overlapping oligopeptides. These oligopeptides can be pentapeptides, heptapeptides, or longer. In general, larger oligopeptides are expected to provide additional insight, but if they are too large the α-helix prediction can become computationally unreasonable. If a protein has N residues, then N-4 pentapeptides must be analyzed to cover the entire protein sequence. The separation of the protein into at least pentapeptides allows the problem to be broken down into a summation of local interactions while still maintaining a central core of three residues to predict the formation of a helical turn.

The rigorous probability values for the helical propensity of each oligopeptide can then be evaluated through detailed atomistic-level modeling using the ECEPP/3 semi-empirical force field (Némethy et al., 1992). The covalent bonds and bond lengths of the oligopeptides are assumed to be fixed at their equilibrium values so each conformation is only a function of its torsional angles. Electrostatic, non-bonded, hydrogen bonded and torsional contributions are combined to yield the total system energy.

After choosing an energy model, in the search for the native peptide conformation it is desirable to identify the oligopeptide conformer with the global minimum energy. Although many approaches have been proposed to solve this problem, few possess deterministic guarantees of the global minimum energy structure. In this approach, each oligopeptide is then analyzed by the use of either modified αBB deterministic branch and bound global optimization techniques (Adjiman et al., 1996, 1997, 1998a,b; Androulakis et al., 1995; Floudas, 2000) or an efficient stochastic genetic algorithm, conformational space annealing (Lee et al., 1998, 1997, 2000; Lee and Scheraga, 1999; Ripoll et al., 1998).

Despite the identification of the minimum energy structure, the selection of the native state cannot be achieved from potential energy calculations alone. The true measure of equilibrium stability of a conformation is the free energy, which must also include the entropic contributions. System information regarding metastable states with the harmonic approximation can determine the accessibility of a given metastable state. This approach requires the generation of a significant ensemble of low potential energy conformations as well as the global minimum potential energy structure. As an added benefit, the method results in occupational probabilities for representative conformations instead of a single conformer.

After generating the low energy ensemble, the entropic contributions to the free energy are calculated at 298 K using the harmonic approximation. This evaluation produces occupational probabilities for each metastable state. By clustering these states according to the backbone torsion angles, an ordered list of conformational propensities is identified for each oligopeptide. An initial helical classification is given to a residue that is a member of the α-helical cluster for more than three consecutive sets of core residues.

However, protein structures in a vacuum environment can be rather different from those in water. Therefore, for oligopeptides containing ionizable residues and exhibiting a high in vacuum probability of helix formation, the effects of solvation and ionization energy must be included to refine the prediction of helical segments. In the current implementation, this entails solving a linearization of the Poisson–Boltzmann equation. Final α-helix propensities are then evaluated by including these additional energy effects to the ensemble classification and again identifying members of the α-helical cluster.

4.2. β-sheet prediction

After the helical segments have been predicted by the described protocol, the remaining residues can be analyzed to predict the location of β-sheets (Klepeis and Floudas, 2003a). The main driving force for this prediction methodology is the role of the hydrophobic collapse in forming β-sheets. The hydrophobic collapse is the process by which the hydrophobic side chains of a protein interact and aggregate, excluding water from the interior of a protein and forming the hydrophobic core. Not all approaches approach the problem this way. The idea of hierarchical folding postulates that a β-sheet nucleates at a hairpin turn and stabilizes itself through a zippering model of hydrogen bond formation (Munoz et al., 1997). Although there has been some debate over the validity of hierarchical folding, recent simulations have supported the alternate view of β-sheet formation through hydrophobic collapse, independent of hydrogen bonding (Bryant et al., 2000; Pande and Rokhsar, 1999; Dinner et al., 1999).

Before modeling the hydrophobic collapse, potential β-strands must be selected. The identification protocol is explicitly designed as an overprediction of the true number of β-strands to produce a superstructure of the possible β-strands of the protein. Then, only those strands selected to participate in the topology by the optimization model are offered as the *beta*-strand prediction.

By properly formulating an Integer Linear Programming (ILP) problem (Floudas, 1995), an objective function representing hydrophobic interaction energy can be maximized over all possible contacts. Three different optimization models have been developed to predict β-sheet topologies in this fashion. Both residue–residue contacts and strand-strand contacts can be used either individually or in a combined model to develop a rank-ordered list of the optimal β-sheet topologies using integer cut constraints.

The first formulation is based on hydrophobic interactions between single residue–residue contacts. For each residue in the hydrophobic set, a hydrophobicity parameter is assigned based on the experimentally derived free energy of amino acid transfer from organic solvents to water (Karplus, 1997; Lesser, 1990; Radzicka and Wolfenden, 1988). The existence of a residue–residue interaction is represented by a binary variable and the hydrophobic contact energies are additive. The objective function for this model maximizes the sum of the hydrophobicities for the β-strands that exist. However, this solution is subject to a number of constraints to allow for realistic topologies. One constraint requires at least one residue–residue contact that is separated by more than 7 amino acids to disallow trivial solutions. A second constraint limits the number of possible hydrophobic contacts for a given residue. Other constraints on the model include the enforcement of symmetric, non-intersecting loops when an antiparallel β-sheet is formed.

A second formulation is based on the idea of a strand-to-strand interaction model. In this model, a hydrophobicity value is assigned to each strand based on the type and count of the hydrophobic amino acids in each strand. The objective function of this model is the maximization of the sum of the strand hydrophobicity values for the strand-to-strand contacts that exist. In addition to the constraints in the residue–residue model, the number of contacts for each strand are limited and specific strand-to-strand contacts can be disallowed.

A final formulation uses both residue–residue contacts as well as strand-to-strand interactions. By combining the objective functions from these two models, both types of contact energies can influence the β-sheet topology prediction. This model contains a representative set of constraints from the first two formulations to again provide feasible β-sheet arrangements. This final formulation is the basis for the current second stage of the ASTRO-FOLD framework.

A globally optimum solution to these ILP problems should be identified to ensure the prediction of the contacts with the maximum hydrophobic interaction energy. These problems are typically solved by a branch-and-bound approach to select the optimal integer solution through a series of Linear Programming relaxations. By including integer cut constraints and iteratively solving the ILP formulation, the best set of competitive β-sheet arrangements can be analyzed as a rank-order list of solutions (Floudas, 1995).

4.3. Restraint derivation and loop modeling

The third step in the ASTRO-FOLD protocol is the development of restraints and the modeling of loop regions. In its unconstrained form, the global optimization of an atomistic-level energy function is an overwhelming challenge. Therefore, attempts to restrict the allowed conformational space of the protein can allow these difficult problems to be solved in manageable amounts of time.

The main feature of protein secondary structure is the resulting patterned formations of local bonding. This knowledge can be used to derive bounds on the dihedral angles and constraints on C^{α}–C^{α} distances. The predicted helical esidues are constrained with an i, $(i+4)$ distance of 5.5 to 6.5 Å where i is a given helical residue and $(i+4)$ is the residue four places away in the primary sequence. This restraint represents the hydrogen bond that results from helix formation. For the residues that are predicted to be helical, the dihedral angles are restricted to $[-90, -40]$ for ϕ and $[-60, -10]$ for ψ.

Similar restraints can be applied to the C^{α}–C^{α} distances in β-sheets. The hydrophobic contacts predicted in the β-sheet optimization models provide tertiary residue–residue contacts to further restrain the system. These contacts are imposed as C^{α}–C^{α} distances between 4.5 and 6.5 Å. The β-sheet dihedral angles are limited to $[-180, -80]$ for ϕ and [80,180] for ψ. The restricted dihedral angle space associated with the secondary structure assignments, along with the imposed contact distances, improve the ability of the tertiary structure prediction algorithm to find the native structure.

Finally, it is desired to model the loop regions between secondary structure elements to develop additional restraints for use in the tertiary structure prediction (Klepeis and Floudas, 2005). Two main methods can be applied to this problem to develop restraints on the dihedral angles and distances. The first method involves an analysis similar to the helical prediction method, where the local interactions are considered through a set of overlapping peptides. The reduced bounds from this analysis are carried forward into subsequent trials with larger loop fragments to finally lead to a simulation of the entire loop. The second method follows a similar approach, but includes longer range interactions by dissecting the distance space over larger fragments of the loop. Simulations of all the domains are conducted and then subsequently combined and used to define appropriate restraints. It is important to note that the simulations of both methods are entirely physics-based, not relying on the fixed position of flanking residues of secondary structure.

4.4. Tertiary structure prediction

After the distance constraints and dihedral angle bounds are included, the goal is to minimize the potential energy of the tertiary structure while satisfying all the constraints. The ASTRO-FOLD approach is a combination of the deterministic αBB global optimization algorithm, a stochastic global optimization algorithm, and a molecular dynamics approach in torsion angle space (Klepeis and Floudas, 2003b,c). The basic formulation is the minimization of the force field energy over torsion angle space, subject to upper and lowering bounding constraints on these angles. Although representing the model in torsion angles increases the model complexity, it significantly reduces the size of the independent variable set.

The use of the αBB global optimization algorithm (Adjiman et al., 1996, 1997, 1998a,b; Androulakis et al., 1995; Floudas, 2000) guarantees convergence to the global minimum solution by a convergence of upper and lower bounds on the potential energy minimum. Upper bounds to this model can be obtained through local minimizations of the original non-convex problem. The addition of separable quadratic terms to the objective and constraint functions produces a convex lower bounding function. With these bounding functions, the problem can be iteratively branched over the variable space, fathoming portions when a region's lower bound rises above the best upper bound.

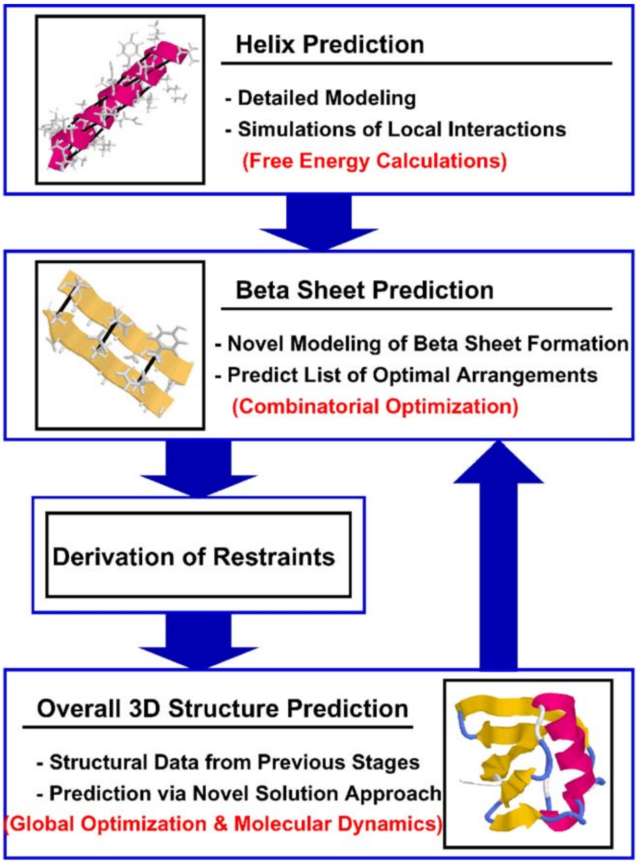
Helix Prediction
- Detailed Modeling
- Simulations of Local Interactions
(Free Energy Calculations)
Beta Sheet Prediction
- Novel Modeling of Beta Sheet Formation
- Predict List of Optimal Arrangements
(Combinatorial Optimization)
Derivation of Restraints
Overall 3D Structure Prediction
- Structural Data from Previous Stages
- Prediction via Novel Solution Approach
(Global Optimization & Molecular Dynamics)

Fig. 2. Overall schematic of ASTRO-FOLD approach for prediction of three-dimensional structure prediction of proteins.

However, as a result of the highly nonlinear force field, this deterministic approach alone is exceptionally difficult. By applying torsion-angle dynamics methods, feasible low energy conformers can be quickly identified, which significantly improve the performance of the αBB method. In addition, the upper bounding approach of the formulation can be augmented by the inclusion of stochastic optimization methods. One such hybrid global optimization method, described as Alternating Hybrids, has been recently introduced (Klepies et al., 2003a,b). It combines the deterministic αBB approach with the stochastic approach of conformational space annealing (Lee et al. 1998, 1997, 2000; Lee and Scheraga, 1999; Ripoll et al., 1998). Conformational space annealing balances genetic algorithm approaches of mutations and crossovers with simulated annealing to identify low energy conformers. In this Alternating Hybrid approach, the search for the native state becomes much more efficient, while still retaining the deterministic guarantees of convergence (Klepeis et al., 2003b).

Although the ASTRO-FOLD framework can be applied to proteins of any size, the detailed energetics and deterministic guarantees of the approach are best-suited for small to medium-sized proteins (up to approximately 200 amino acids in length). The methodology has been successfully applied to a varied set of proteins throughout this range (Klepeis and Floudas, 2003c). A diagram of the current ASTRO-FOLD protocol is presented in Fig. 2 and a recent double blind prediction study is reported in Klepeis et al. (2005).

5. Force fields

Protein structure prediction is one of the most important and difficult problems in computational structural biology. As discussed earlier, different approaches have been developed to address this problem. Various components of the protein folding problem (e.g., fold recognition, ab initio prediction, comparative modelling and de novo design) make use of a force field. In the process of structure prediction, sometimes it is required to select the native structure of a protein from a pool of non-native structures. Force fields are used to select the native structure. It uses various interactions occurring at the atomic level to calculate the energy of the conformer. If the energy function includes every type of interaction present in a detailed atomic model of a protein then it is called a true effective energy function. These true effective energy functions can be obtained by applying basic laws of physics at the atomic level of a protein. However, increased computational effort is needed because atomistic level formulation requires consideration of energetics between all pairs of atoms and the number of pairs increases rapidly as the chain length increases.

Some of the semi-empirical force fields commonly used are CHARMM (MacKerell Jr. et al., 1998), AMBER (Cornell et al., 1995), ECEPP (Momany et al., 1975), ECEPP/3 (Némethy et al., 1992), and GROMOS (Scott et al., 1997). It has been pointed out that even these types of potentials are not very effective in discriminating native and non-native structures (Novotny et al., 1984; Wang et al., 1995). Hence, a lot of effort has been invested in finding a simplified protein potential which is capable of differentiating native and non-native proteins without heavily increasing the computational load. These simplified force fields do not demand huge computational resources and use a coarse-grained description of the protein structure. This section discusses recent developments in this field. Publications by Shakhnovich (1998), Levitt et al. (1997) and Hao and Scheraga (1999) can be referred for review of work done in this field. In this presentation, specific contributions by each group have been given.

Maiorov and Crippen (1992) introduced a linear programming concept in determining the force field. They calculated the potential by accommodating a large number of parameters corresponding to interaction energy between two amino acids, and forcing the native structure to have the lowest energy among other alternatives. This model used a "continuous contact function" instead of "square-well function" because they suspected that small changes in interatomic distances between two homologous structures might lead to unwanted changes in contact energy. This model was tested on 10,000 decoys of 37 proteins and was found to perform better than the work by Hendlich et al. (1990). Ohkubo and Crippen (2000) developed a pairwise potential, trained it for one protein and showed that this kind of potential can be used to allocate the lowest energy to the native conformation. This result was in contradiction with work done by Vendruscolo and Domany (1998). In one of their recent works (Crippen, 2001) they described various requirements for a good force field in terms of stability and sensitivity of native structure over its conformers and developed a new force field based on non-hydrogen atoms that can satisfy most of these requirements.

Vendruscolo and Domany (1998) used a contact map representation and developed a C^{α} distance dependent force field. They proved that it is not possible to calculate a set of interaction energy parameters that can be used to assign a higher energy to all non-native structures than to the native structure of any protein. In this formulation they used a contact definition based on C^{α} carbon atoms and the cutoff distance was taken as 8.4 Å. This problem can be overcome by using a more accurate definition of contact [for example distance dependent definition] as shown by Loose et al. (2004) and Tobi and Elber (2000). Clementi et al. (1999) also supported the work by Vendruscolo and Domany (1998). They designed protein-like heteropolymer sequences and tried to approximate the Lennard–Jones potential by a pairwise contact potential. They also showed that there is no pairwise contact potential that can satisfy the requirements of a force field. In all these works, Domany and coworkers used a specific definition of "contact" and based their results on the basis of this definition. In another published work Vendruscolo et al. (2000) studied the effect of various parameters [for example, definition of contact, cut-off length and number of proteins used]. After studying the effect of these parameters they reached the same conclusion.

The idea of using Boltzmann distribution to find knowledge-based force field was first introduced by Tanaka and Scheraga (1976). This assumption has been justified by various authors: Bryant and Lawrence (1991), Finkelstein et al. (1995), even though the definition of reference state in Boltzmann distribution is critical. Jernigan and Bahar (1996) reviewed the choice of this reference state. They represented a residue by two interaction sites, one on the backbone and another on the amino acid side group (Bahar and Jernigan, 1997). One of the important findings of this work was that hydrophillic interactions are important up to a distance range of 4 Å and with an increase in inter-residue distance [less than 6.4 Å], the effect of hydrophobic interaction also increases. These experimental results are very useful and can be used as constraints when using an optimization approach to develop a force field (Loose et al., 2004). Miyazawa and Jernigan (1999) further improved their model by adding terms corresponding to secondary structure potential, tertiary structure potential and repulsive packing potential while calculating total conformational energy. It was shown by a gapless threading experiment that this potential can be used in both fold and sequence recognition.

Significant contributions in this field were introduced by Scheraga and coworkers. They developed a united residue representation (UNRES) of polypeptide chain (Liwo et al., 1997a,b, 1998). In this model the polypeptide chain is modelled as a sequence of C^α carbon atoms each connected by a hypothetical bond. The center of this hypothetical bond along with the center of the side chain is taken as interaction centers. Energetic calculations are done by considering possible interactions between these centers. While using this kind of representation, details of other atoms are lost. Liwo et al. (1997c) used a cooperative term (as used by Godzik et al., 1993) to include the multi-body effect. Multi-body effect considers the interaction between more than two atoms. This is more accurate than two body effect as details of other atoms are also used while calculating interaction energies. Liwo et al. (1997b) parameterized various functional forms of interactions. They also calculated appropriate weight terms by minimizing the z-scores. Liwo et al. (2001) further improved the UNRES model by introducing some additional terms (e.g., corresponding to hydrogen bonding) which enabled the model to predict the beta sheets more efficiently. Pillardy et al. (2001) developed three different force fields using the UNRES model. These force fields were developed by calculating weight factors by considering only α-helical, β-sheet, α–β proteins, respectively. If the model is trained specifically for either α-helical or β-sheet protein, then it might be possible to predict native structures of proteins with only α-helical and β-sheets more effectively. In one of their most recent work, Liwo et al. (2004) re-parameterized the backbone-electrostatic and multi-body contributions terms of the UNRES model to make it more efficient.

Tobi et al. (2000) developed a pairwise distance dependent model using linear programming. In this work, inter-residue distance was divided in 7 bins and an energy parameter corresponding to each bin and amino-acid interaction was calculated using optimization techniques. MONSSTER (Skolnick et al., 1997b) was used to generate 4,299,167 decoys for 75 proteins. MONSSTER (MOdeling of New Structures from Secondary and TErtiary Restraints) is a method for folding globular proteins using small number of distance restraints. Using this large number of decoys they concluded that there does not exist any pairwise function with 1 Åresolution that can fold a protein to its native structure. Tobi and Elber (2000) divided the inter-residue distance in 13 bins. They also proposed that further refinement of the distance bin will not serve any purpose because the size of the amino acid is already a few angstroms (Tobi and Elber, 2000). The results of this formulation were in agreement with the results of Bahar and Jernigan (1997). When using a linear programming approach, it is possible to get an infeasible solution. For the cases of infeasible solution, Meller et al. (2002) developed a maximum feasibility guideline that can be used to find the best possible potential using a subset of data. This heuristic finds application in cases involving large data space (Meller et al., 2002). This guideline was used by Loose et al. (2004) to formulate an iterative constraint-dropping scheme to identify the largest feasible subset of constraints.

While using the Boltzmann distribution to derive the potential function, a quasichemical approximation is implicitly assumed. This approximation assumes that amino acid residues are disconnected units. Jernigan and Bahar (1996) questioned this assumption and outlined the possible error introduced by using this assumption. Skolnick et al. (1997a) assessed the validity of this approximation by including the effect of chain connectivity, and the presence of secondary structure. This force field used a contact based potential with a cutoff distance of 4.5 Å. After solving this model they concluded that these considerations do not have any effect on the derived potential (Zhang and Skolnick, 1998). In another work they developed a heavy atom distance dependent force field (Lu and Skolnick, 2001). Using heavy atoms instead of C^α atoms increased the number of residue centers from 20 to 167. Furthermore, the distance between these interaction centers was divided into 14 bins. A Boltzmann distribution with three different scales of reference state was used to calculate the contact frequency in reference state. After solving this model, an improvement of 4 units of z-score was observed when tested on a gapless threading set.

In an attempt to determine an accurate force field Samudrala and Moult (1998) employed an all-atom approach. Instead of using some representative interaction centers, they used conditional probability based, all-atom description to determine the force field. This force field was tested on various decoy sets and it performed quite well on a variety of decoys. Similar to the work of Samudrala and Moult (1998), Zhou and Zhou (2002) also developed an all-atom, distance dependent force field. In particular, they proposed a new reference state called DFIRE (distance-scaled, finite ideal-gas reference). They used this reference state in their calculation of all-atom based potential and showed that this force field is much better than previously developed, residue-specific force field both in terms of discrimination capacity and z-scores (Zhou and Zhou, 2002). Zhang et al. (2004) presented a simplified version of this all-atom based force field (Zhou and Zhou, 2002) by using the center of mass of side chains. Upon testing they found that despite this simplification the performance of the force field does not change by much.

A distance dependent C^{α}–C^{α} based force field has been recently developed by Loose et al. (2004). They introduced a linear programming approach which combines information from experimental observation as explicit constraints. A high quality decoy set was used to train this model. Decoys generated by DYANA (Güntert et al., 1997) were minimized by TINKER (Ponder and Richards, 1987a). DYANA uses sequence and secondary structure information of a protein to minimize the energy of the structure. After minimization, a molecular dynamics simulation in torsion angle space is used to change the shape of the protein and Van der Waal contact energy is minimized further to generate low-energy decoys. These high quality decoys were generated for 758 proteins and approximately 108,900 decoys were used to train this model. A number of additional constraints were added to incorporate the results obtained from different experiments (Jernigan and Bahar, 1996). The resulting force field's performance was tested on a set of 151 new proteins and this force field outperformed the TE-13 force field (Tobi and Elber, 2000). These decoys can be found in http://titan.princeton.edu/Decoys.

Evaluation and testing of force field is also an important step in development of force fields. High quality decoys are needed to test the effectiveness of a force field. Tsai et al. (2003) developed an improved decoy set of 1400 structures of 78 dissimilar proteins. Another set of decoys that can be used to test force fields can be found on the PROSTAR website [http://prostar.carb.nist.gov/].

Important applications of force fields are also in the area of de novo protein design. Very often the objective function to minimize in the de novo protein design problem is energy, and in order to calculate energy an energy function or force field is required. Researchers working in the area of de novo protein design have developed simpler force fields using the molecular mechanics force fields as a foundation and applied them for sequence selection (Gordon et al., 1999).

These mean force fields for protein design still contain elements accounting for van der Waals force, electrostatics, solvation, and hydrogen bonding. Van der Waals force is usually described by the equation for Lennard–Jones 6–12 potential. Unlike van der Waals interactions, electrostatics, solvation, and hydrogen bonding for proteins existing in an aqueous environment are not well explained by molecular mechanics force fields (Edinger et al., 1997) and hence more parametric fitting with empirical data is usually necessary. Electrostatic interactions for protein design are difficult to model because they are highly dependent on the local environment of the residues and thus cannot be treated generally (Pokala and Handel, 2001). Solvation and electrostatics are closely knitted because solvent molecules can either interact directly with the charged or polar residues on the protein surface, or affect electrostatics indirectly by shielding local charges. Currently there are two main classes of electrostatics/solvation models. One class applies reduced or distance-dependent dielectric constants and a surface area-dependent term that promotes the exposure of polar groups to the solvent. The other class employs finite-difference approximations to the Poisson–Boltzmann continuum dielectric model (Honig et al., 1993), which assumes that a protein can be treated as a low-dielectric charged object placed in a high-dielectric solvent medium. The latter class is in general too computationally expensive for de novo protein design, particularly for full sequence optimization. The former class is simpler and thus more implementable but it gives coarser results. However, parametrization can improve the results. Wisz and Hellinga (2003) introduced into the dielectric constant dependence on protein geometry, local environment, and the particular types of amino acids under interactions. They obtained results comparable to those from the Poisson–Boltzmann continuum dielectric model. As for hydrogen bonding, it demands accurate simulation as it plays a key role in protein structure stabilization and protein specificity. Like electrostatics and solvation, parametrization efforts on hydrogen bonding modelling are commonly found. For example, Kortemme et al. (2003) had parameterized hydrogen bonds with three different angles and one distance to make it orientation-dependent, and proved that their model was superior to van der Waals like models for hydrogen bonding.

6. De novo protein design

There have been considerable successes in the development of computational algorithms for protein design during the past decade. At the turn of the 1990s, some de novo protein design efforts turned out to be futile as either the target fold was not achieved (Betz et al., 1993) or the engineered protein had a different quaternary structure than expected (Lovejoy et al., 1993). These failures were thought to be caused by the relatively qualitative hierarchic approach adopted on protein design at that time (Street and Mayo, 1999). Then there came several successful computational protein design attempts, most of which having focused on the protein cores. The reason why protein core was picked instead of the surface or boundary is that protein folding is primarily driven by hydrophobic collapse, and thus a good core will enable a well-folded and stable structure for the de novo designed protein (Dill, 1990). Several research groups in the field have applied *in silico* methods to design the hydrophobic cores, with the novel sequences being validated with experimental data (Richards and Hellinga, 1994; Desjarlais and Handel, 1995; Dahiyat and Mayo, 1996). As time went on, the focus of study has been broadened to include surface residues. This obviously required more accurate energy functions or force fields. Recently, *in silico* protein design has encompassed rendering novel functions on templates originally lacking those properties, modifying existing functions, and increasing protein stability or specificity. Beyond any doubt, intense research activities are ongoing in the field, the potential of which is simply enormous.

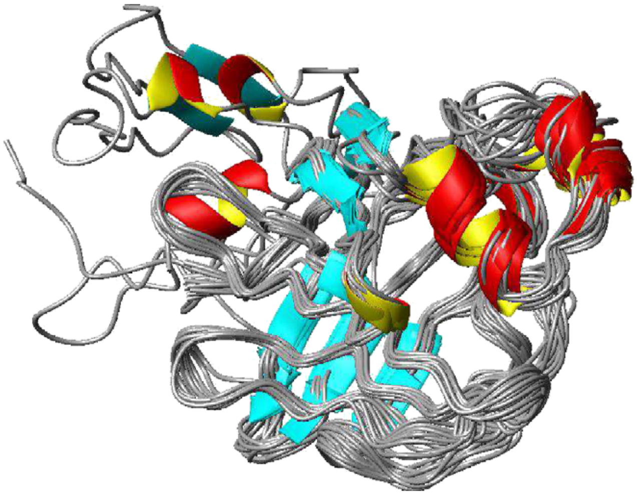

Fig. 3. Template flexibility as illustrated by the overlapping of the 20 available NMR structures of Apo Intestinal Fatty Acid-Binding Protein (PDB code:1AEL).

6.1. Template flexibility

Many computational protein design efforts were based on the premise that the template or backbone was rigid and thus the three-dimensional coordinates of all the atoms on the template were fixed. This assumption was first proposed by Ponder and Richards (1987b), and was appealing because it greatly reduced the search space and thus the time required to converge to a solution for the minimum energy sequence, regardless of the kind of search method employed. However, the assumption was also highly questionable, as the template was commonly known to exhibit flexibility in real situations as shown in Fig. 3. Protein backbones had been observed to allow residues that would not have been permissable had the backbone been fixed (Lim et al., 1994). In the Protein Data Bank, there exist numerous examples of proteins which exhibit multiple NMR structures. Though commonly assumed as rigid bodies as a first approximation, the secondary structures of α-helices and β-sheets actually display some twisting and bending in the protein fold, and Emberly et al. (2003, 2004) had applied principal component analysis of database protein structures to quantify the degree and modes of their flexibility.

The Mayo group (Su and Mayo, 1997; Ross et al., 2001) had claimed that their ORBIT (Optimization of Rotamers By Iterative Techniques) computational protein design process was robust against 15 per cent change in the backbone. Nevertheless, they found out on a later case study on T4 lysozyme that core repacking to stabilize the fold was difficult to achieve without considering a flexible template (Mooers et al., 2003). Therefore, to ensure that good sequence solutions are not rejected, it is more desirable to assume backbone flexibility in de novo protein design.

Researchers have formulated several methods to incorporate template variability. First, backbone flexibility can simply be modeled by using a smaller atomic radii in the van der Waals potential. One common practice has been to scale down the radii by five to ten per cent (Desjarlais and Handel, 1995; Kuhlman and Baker, 2000) and thus permitting slight overlaps between atoms due to backbone movements. Key disadvantages of this simple approach include overestimation of the attractive forces and also the possibility of hydrophobic core overpacking.

Another way to allow for backbone flexibility is through considering a discrete set of templates by using genetic algorithms and Monte Carlo sampling. This is the approach adopted by both Desjarlais and Handel (1999) and Kraemer-Pecore et al. (2003). Under this approach an ensemble of related backbone conformations close to the template are generated at random. Then a sequence will be designed for each of them under the rigid backbone assumption, and finally the backbone-sequence combination with the lowest energy will be selected. For symmetric proteins backbone structure can actually be modeled by parametric fitting and this will enhance computational efficiency. Backbone parametrization is performed by deriving quantitative relations between structural parameters and overall protein secondary structures. For instance, the number of strands and a measure of stagger in a β-sheet can be related to the general shape and the twisting and coiling of the β-sheet (Murzin et al., 1994a,b). However, the vast majority of protein structures are non-symmetric which make this parametric approach infeasible. Su and Mayo (1997) overcame this difficulty by treating α-helices and β-sheets as rigid bodies and designing sequences for several template variations of the protein Gβ1. Farinas and Regan (1998) considered a discrete set of templates when they designed the metal binding sites in Gβ1, and they identified varied residue positions that would have been missed if average three-dimensional coordinates had been used for calculations. Harbury et al. (1998) incorporated template flexibility through an algebraic parameterization of the backbone when they designed a family of α-helical bundle proteins with right-handed superhelical twist. They were able to achieve a root mean square coordinate deviation between the predicted structure and the actual structure of the de novo designed protein of around 0.2 Å. Larson et al. (2002) considered backbone flexibility by designing protein to structural ensembles, the optimal number of which being determined by the sequence entropy. They generated hundreds of thousands of sequences for 253 naturally occurring proteins, and with homology search they claimed the diverse sequence space obtained was of high quality.

One natural approach to incorporate backbone flexibility is to allow for variability in each position in the template. The deterministic *in silico* sequence selection method recently proposed by Klepeis et al. (2003c, 2004) using integer linear optimization technique takes into account template flexibility via the introduction of a distance dependent force field in the sequence selection stage. Pairwise amino acid interaction potential, which depends on both the types of the two amino acids and the distance between them, were used to calculate the total energy of a sequence. Instead of being a continuous function, the dependence of the interaction potential on distance is discretized into bins. With typical bin sizes of 0.5 to 1 Å, the overall protein design model Klepeis et al. (2004) developed implicitly incorporated backbone movements of roughly the same order of magnitude.

6.2. Amino acid sequence search methods

A common target for de novo protein design is the ability to perform full sequence search for a 100-residue protein (Street and Mayo, 1999; Saven, 2002; Pokala and Handel, 2001). Considering that there are 20 naturally occurring amino acids for each position, the combinatorial complexity of the problem amounts to 20^{100} or 10^{130}. This often proves to be an insurmountable task even with the computational power of modern computers.

One simple and popular approach to reduce the search space immediately is to classify the protein residues into categories according to their environment, namely core amino acids, surface amino acids, and boundary amino acids. Since surface residues are those exposed to the aqueous environment, they will most likely be hydrophilic amino acids. In contrast, the protein core is concealed from water and thus is composed of mainly hydrophobic amino acids. The boundary residues have to be selected from the full range of 20 amino acids as these positions can be either hydrophilic or hydrophobic. Hecht et al. (2004) applied the binary pattern of polar and non-polar amino acids to design and synthesize their protein combinatorial libraries in an attempt to search for protein-like properties. They used the binary code to characterize secondary structures: for α-helices, the sequence periodicity of polar and non-polar amino acids must roughly match the structural repear of 3.6 residue per turn; for β-sheets, polar and non-polar amino acids have to alternate in every other position (Hecht et al., 2004).

Focusing on a subset of residues for each position, if possible, allows for the use of deterministic methods in search for the sequence with the lowest energy objective. The self-consistent mean field (SCMF) (Lee, 1994) and dead-end-elimination (DEE) (Desmet et al., 1992) are both good examples of deterministic methods. SCMF tests an ensemble of amino acid/rotamer combinations at each position of a fixed template, with each rotamer in the ensemble given the same Boltzmann probability. The rotamer Boltzmann probabilities for all other positions are then computed to obtain a weighted average energy, which is used to recalculate the Boltzmann probability for each rotamer for each position. Thus, the process is iterative and will terminate when the Boltzmann probability converges to a certain value. The main disadvantage of SCMF is that though deterministic in nature, it does not guarantee to yield a global minimum in energy (Lee, 1994). In contrast, DEE assures the convergence to a globally optimal solution consistently. DEE operates on the systematic elimination of rotamers that are not allowed to be parts of the sequence with the lowest energy. The energy function in DEE is written in the form of a sum of individual term (rotamer–template) and pairwise term (rotamer–rotamer). The Mayo group has pioneered the development of DEE and has applied the method to design a variety of proteins (Malakauskas and Mayo, 1998; Strop and Mayo, 1999; Shimaoka et al., 2000; Bolon and Mayo, 2001; Mooers et al., 2003). Goldstein (1994) relaxed the rotamer elimination criterion in DEE to address problems of bigger size. Pierce et al. (2000) introduced Split DEE which split the conformational space into partitions and thus eliminated the dead-ending rotamers more efficiently. On the other hand, Looger and Hellinga (2001) introduced generalized DEE by ranking the energy of rotamer clusters instead of individual rotamers and increased the ability of the algorithm to deal with higher levels of combinatorial complexity. Further revisions and improvements on DEE had been performed by Wernisch et al. (2000) and Gordon et al. (2003). The key limitations imposed on the SCMF and DEE are (i) the backbone/template is fixed, and (ii) sequence search is restricted to discrete set of rotamers.

Klepeis et al. (2003c, 2004) introduced a novel two-stage protein design approach which first solved for the lowest energy sequences in the form of a rank ordered list using an ILP formulation, and then validated fold specificity by calculating the ensemble probabilities of those sequences obtained in the previous stage. This approach allows naturally for template flexibility and rigorously enumerates the sequence space. They have tested their predictions against experimental data on the inhibitory activity of compstatin, a 13-residue peptide, and have obtained 16-fold improvements over the parent peptide (Klepeis et al. 2003c, 2004).

The protein design problem has been proved to be NP-hard (Pierce and Winfree, 2002), which means the time required to solve the problem varies exponentially according to n^m, where n is the average number of amino acids to be considered per position and m is the number of residues. Hence as the protein becomes big enough, deterministic methods may reach a plateau, and this is when stochastic methods come into play. Monte Carlo methods and genetic algorithms are the most commonly used stochastic methods for de novo protein design. In Monte Carlo methods, a mutation is performed at a certain position in the sequence and the Boltzmann probability calculated from the energies before and after the mutation, as well as temperature is compared to a random number. The mutation is allowed if the Boltzmann probability is higher than the random number, and rejected otherwise. The Baker group (Kuhlman et al., 2002; Dantas et al., 2003; Kuhlman et al., 2003)'s protein design computer program, RosettaDesign, applied Monte Carlo optimization algorithms. In completely redesigning nine globular proteins, RosettaDesign yielded sequences of 70–80% identity as the final results of energy optimization when multiple runs were started with different random sequences (Dantas et al., 2003). Originated in genetics and evolution, genetic algorithms generate a multitude of random amino acid sequences and exchange for a fixed template. Sequences with low energies form hybrids with other sequences while those with high energies are eliminated in an iterative process which only terminates when a converged solution is attained. Desjarlais and Handel (1999) have applied a two-stage combination of Monte Carlo and genetic algorithms to design the hydrophobic core of protein 434cro. Both Monte Carlo methods and genetic algorithms can search larger combinatorial space compared to deterministic methods, but they share the common disadvantage of lacking consistency in finding the global minimum in energy.

Lastly, it should be noted that instead of searching for the whole sequences with the lowest energies, de novo protein design has been performed by assigning probability to an amino acid for each position in a sequence that will fold into the three-dimensional target structure. The set of site-specific amino acid probabilities obtained at the end actually represents the sequence with the maximum entropy subject to all of the constraints imposed (Zou and Saven, 2000; Kono and Saven, 2001; Park et al., 2004). This statistical computationally assisted design strategy (*scads*) has been employed to characterize the structure and functions of membrane protein KcsA and to enhance the catalytic activity of a protein with dinuclear metal center (Park et al., 2004). It has also been used to calculate the identity probabilities of the varied positions in the immunoglobulin light chain-binding domain of protein L (Kono and Saven, 2001). *Scads* serves as a useful framework for interpreting and designing protein combinatorial libraries, as it provides clues about the regions of the sequence space that are most likely to produce well-folded structures (Hecht et al., 2004).

6.3. Successes and prospects

So far there have been numerous examples of full sequences designed “from scratch” that were confirmed to fold into the target three-dimensional structures by experimental data (Walsh et al., 1999; Bryson et al., 1998). The zinc-finger protein designed by (Dahiyat and Mayo, 1997) was the first one to appear. Recently, Kraemer-Pecore et al. (2001) also performed a full-sequence design on the WW motif, a β-sheet protein and verified with spectroscopic data that it had a structure similar to that of the natural protein. The successes on achieving target folds should be attributed to the quantitative tools researchers use for confirming fold specificity. Emberly et al. (2002a), Emberly et al. (2002b), Li et al. (2002), and Miller et al. (2002) used what they called designability, which is essentially the number of sequences that have the desired structure as their lowest energy state, to ensure the target fold. Saven (2001, 2003) applied the energy landscape theory of protein folding to develop the foldability criterion for confirming sequence-structure compatibility. Koehl and Levitt (1999a) explored both the sequence space and the conformation space to generate sequences compatible with the template (the so-called “design in” procedure) and incompatible with the competing non-native folds (the so-called “design out” procedure) and redesigned the B1 domain of protein G, the lambda repressor, and the sperm whale myoglobin with their respective native structures. Their novel approach for protein design guaranteed the specificity of the designed sequence for the template by fixing the amino acid composition, and they proved this new procedure converged in sequence space (Koehl and Levitt, 1999b).

The ultimate goal of computational protein design is of course not just to achieve the desired structure but also to render specific functions or properties to the novel protein. In the latter respect, research efforts were found to be multi-faceted computational protein design had been applied to design protein–protein interaction specificity, enhance protein stability, confer brand-new metal binding centers onto proteins originally lacking those moieties, create proteins that fold faster than the mother sequences, and predict sequence mutations that restrict proteins in certain conformations.

Using computational methods, Wilson et al. (1991) screened amino acids for the active site on α-lytic protease to improve specificity toward the substrate. The engineered enzyme had an over 200-fold preference for substrates with Leu at the P1 position over those with Ile at the same position and met the peptide design objective. Klepeis et al. (2003c, 2004) optimized six residue positions out of a total of thirteen positions on compstatin, both a peptide that inhibits complement activation and a strong candidate for being a pharmaceutical. Sequences from computational results were synthesized in the laboratory and the one with the strongest inhibitory activity was found to be 16-fold more potent than the parent peptide compstatin. Reports from Ghirlanda et al. (1998) on their design of a two-helix receptor that binds to the calmodulin binding domain (CBD) of calcineurin pointed out the importance of negative design for achieving high protein–protein interaction specificity (Pokala and Handel, 2001). Negative design refers to including and employing the thermodynamics information about all unfolded states of the target protein during the de novo design process. The concept originated in the free energy landscape theory of protein folding which states that proteins traverse a smooth funnel-shaped energy landscape during folding and the minimum energy corresponds to the folded conformation. For a de novo designed protein to have a stable structure, there must be a significant energy difference between the folded and unfolded states, normally taken to be greater than the energy fluctuations among the unfolded states. This is the reason why thermodynamics of non-native competing structures have to be taken into account if high specificity is to be attained.

Another major application for de novo protein design lies in promoting stability of the target protein. One easy way to achieve this can be to increase the hydrophobic area buried (Malakauskas and Mayo, 1998). It was found that designed proteins with more hydrophobic amino acids than the parent proteins from which they were derived would usually have higher stability (Kuhlman and Baker, 2004). This can be a simple rule of thumb that researchers performing protein design should keep in mind. Increasing hydrophobic residues can be performed by either switching partially buried hydrophilic or charged residues to non-polar residues, or packing more hydrophobic moieties in the core.

As for conferring novel metal bind sites onto a template, Richards and Hellinga (1991) had proposed the DEZYMER program, a strategy which first identified the catalytic functional groups that will catalyze the desired reaction and then relocated those groups from the mother protein to the best positions in the de novo designed protein. DEZYMER program had been applied to create zinc, iron sulphide, and copper binding sites in thioredoxin, a protein that normally does not bind to metal ions at all (Richards et al., 1991). Benson et al. (2000) had also designed metalloenzymes *in silico* to imitate the catalytic site in superoxide dismutase (SOD). They correlated catalytic activity with parameters like location and three-dimensional structure of the environment of the binding site. This implies de novo protein design can play a vital role in understanding redox reactions in protein. Other successes on de novo protein design from the Hellinga group were reported by Dwyer et al. (2003, 2004), Looger et al. (2003) and Allert et al. (2004).

Proteins de novo designed to be consistent with the three-dimensional target with minimum energy interactions were often found to fold very fast, with folding times of 1 to 50 μs (Gillespie et al., 2003; Zhu et al., 2003). This agrees with the observation of Watters and Baker (2004) about the src SH3 domain and that of Kuhlman and Baker (2004) about seven computer-generated proteins most of the times the engineered proteins folded faster than the wild type proteins. The reason why de novo designed proteins have faster folding kinetics compared to the natural ones is not exactly clear, but the common consensus is that evolution has not operated on protein folding rates but rather on stability of the folds. However, this interesting phenomenon indicates that *in silico* protein design can have broad applications on investigating protein folding kinetics. Kuhlman and Baker (2004) utilized computational protein design techniques and managed to alter the folding pathways of the IgG-binding domains of protein G and protein L.

Finally, proteins are macromolecules and their high degree of freedom in motions allow them to have multiple conformations. In addition, very often some conformations are preferred over the others from the protein engineering point of view. De novo protein design had been used by researchers to lock proteins into certain useful conformations. For instance, Shimaoka et al. (2000) had computed variants of the integrin I, a cell-surface adhesion receptor that interacts with the complement component iC3b, to enforce the protein to adopt either the open or closed conformation. This conclusion was drawn based on the observation that variants designed to mimic the open conformation show higher binding affinities with the ligand than the variants designed to mimic the closed conformation. In addition, Kraemer-Pecore et al. (2001) also claimed success in restricting the amino-terminal domain of calmodulin to its calcium saturated closed form.

It is expected that de novo designed proteins with few hundred residues will soon be in place (Pokala and Handel, 2001). They will certainly be able to fold into the target structure, carrying all the desired properties and functions prescribed at the design stage. However, before that actually happens, further efforts definitely have to be put into the development of better force fields or scoring functions, more powerful and accurate search methods, and faster and more systematic ways of screening against experimental data. Recently, active research is underway to incorporate unnatural amino acids into computational protein design (Sia and Kim, 2001; Tang et al., 2001; Mallik et al., 2005). This brings the de novo protein problem to a higher level of complexity but makes it more fascinating, and will certainly trigger a new wave of computational algorithms for solving the problem.

Acknowledgements

CAF gratefully acknowledges financial support from the National Science Foundation and the National Institutes of Health (R01 GM52032; R24 GM069736). MM gratefully acknowledges funding by a Deutsche Forschungsgemeinschaft research fellowship (MO 1086/1-1).

References

Adjiman, C.S., Androulakis, I.P., Maranas, C.D., Floudas, C.A., 1996. A global optimization method, αBB, for process design. Computers and Chemical Engineering 20, S419–S424.

Adjiman, C.S., Androulakis, I.P., Floudas, C.A., 1997. Global optimization of MINLP problems in process synthesis and design. Computers and Chemical Engineering 21, S445–S450.

Adjiman, C.S., Androulakis, I.P., Floudas, C.A., 1998a. A global optimization method for general twice-differentiable NLPs—II. Implementation and computational results. Computers and Chemical Engineering 22, 1159–1179.

Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A., 1998b. A global optimization method for general twice-differentiable NLPs—I. Theoretical advances. Computers and Chemical Engineering 22, 1137–1158.

Allert, M., Rizk, S., Looger, L.L., Hellinga, H.W., 2004. Computational design of receptors for an organophosphate surrogate of the nerve agent soman. Proceedings of the National Academy of Sciences of the United States of America 101, 7907–7912.

Aloy, P., Stark, A., Hadley, C., Russell, R.B., 2003. Predictions without templates: new folds secondary structure, and contacts in CASP5. Proteins: Structure, Function, and Bioinformatics 53, 436–456.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215 (3), 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402.

An, Y., Friesner, R.A., 2002. A novel fold recognition method using composite predicted secondary structures. Proteins: Structure, Function, and Bioinformatics 48, 352–366.

Androulakis, I.P., Maranas, C.D., Floudas, C.A., 1995. αBB: a global optimization method for general constrained nonconvex problems. Journal of Global Optimization 7, 337–363.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181 (4096), 223–230.

Bahar, I., Jernigan, R.L., 1997. Inter-residue potential in globular proteins and the dominance of highly specific hydrophillic interactions at close separation. Journal of Molecular Biology 266, 195–214.

Benson, D.E., Wisz, M.S., Hellinga, H.W., 2000. Rational design of nascent metalloenzymes. Proceedings of the National Academy of Sciences of the United States of America 97, 6292–6297.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Research 235–242.

Bertsekas, D.P., 1995. Dynamic Programming and Optimal Control—I. Athena Scientific,

Betz, S.F., Raleigh, D.P., DeGrado, W.F., 1993. De novo protein design: from molten globules to native-like states. Current Opinion in Structural Biology 3, 601–610.

Bolon, D.N., Mayo, S.L., 2001. Enzyme-like proteins by computational design. Proceedings of the National Academy of Sciences of the United States of America 98, 14274–14279.

Bowie, J.U., Lüthy, R., Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253 (5016), 164–170.

Bradley, P., Chivian, D., Meiler, J., Misura, K.M.S., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E.M., Baker, D., 2003. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins: Structure, Function, and Bioinformatics 53, 457–468.

Bryant, S.H., Lawrence, C.E., 1991. The frequency of ion-pair substructures in proteins is quantitaively related to electrostatic potential. A statistical model for nonbonded interactions. Proteins: Structure, Function, and Bioinformatics 9, 108–119.

Bryant, Z., Pande, V.S., Rokhsar, D.S., 2000. Mechanical unfolding of a beta-hairpin using molecular dynamics. Biophysical Journal 78, 584–589.

Bryson, J.W., Desjarlais, J.R., Handel, T.M., DeGrado, W.F., 1998. From coiled coils to small globular proteins: design of a native-like three-helix bundle. Protein Science 7, 1404–1414.

Chothia, C., 1992. One thousand families for the molecular biologist. Nature 357, 543–544.

Clementi, C., Vendruscolo, M., Maritan, A., Domany, E., 1999. Folding Lennard–Jones proteins by a contact potential. Proteins: Structure, Function, and Bioinformatics 37, 544–553.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz Jr., K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., Kollman, P.A., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society 117, 5179–5197.

Crippen, G.M., 2001. Constructing smooth potential functions for protein folding. Journal of Molecular Graphics and Modelling 19, 87–93.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M., Barton, G.J., 1998. JPred: a consensus secondary structure prediction server. Bioinformatics 14 (10), 892–893.

Czaplewski, C., Liwo, A., Pillardy, J., Oldziej, S., Scheraga, H.A., 2004a. Improved conformational space annealing method to treat beta-structure with the UNRES force-field and to enhance scalability of parallel implementation. Polymer 45, 677–686.

Czaplewski, C., Oldziej, S., Liwo, A., Scheraga, H.A., 2004b. Prediction of the structures of proteins with the UNRES force field, including dynamic formation and breaking of disulfide bonds. Protein Engineering Design and Selection 17, 29–36.

Dahiyat, B.I., Mayo, S.L., 1996. Protein design automation. Protein Science 5, 895–903.

Dahiyat, B.I., Mayo, S.L., 1997. De novo protein design: fully automated sequence selection. Science 278, 82–87.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., Baker, D., 2003. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. Journal of Molecular Biology 332, 449–460.

de Bakker, P.I.W., DePristo, M.A., Burke, D.F., Blundell, T.L., 2003. Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized Born solvation model. Proteins: Structure, Function, and Bioinformatics 51, 21–40.

Deane, C.M., Blundell, T.L., 2001. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Science 10, 599–612.

DePristo, M.A., de Bakker, P.I.W., Lovell, S.C., Blundell, T.L., 2003. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins: Structure, Function, and Bioinformatics 51, 41–55.

Desjarlais, J.R., Handel, T.M., 1995. De novo design of the hydrophobic cores of proteins. Protein Science 4, 2006–2018.

Desjarlais, J.R., Handel, T.M., 1999. Side-chain and backbone flexibiity in protein core design. Journal of Molecular Biology 290, 305–318.

Desmet, J., Maeyer, M., Hazes, B., Lasters, I., 1992. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356, 539–542.

Dill, K.A., 1990. Dominant forces in protein folding. Biochemistry 29, 7133–7155.

Dinner, A.R., Lazaridis, T., Karplus, M., 1999. Understanding beta-hairpin formation. Proceedings of the National Academy of Sciences of the United States of America 96, 9068–9073.

Drexler, K.E., 1981. Molecular engineering: an approach to the development of general capabilities for molecular manipulation. Proceedings of the National Academy of Sciences of the United States of America 78 (9), 5275–5278.

Dwyer, M.A., Looger, L.L., Hellinga, H.W., 2003. Computational design of a Zn2+ receptor that controls bacterial gene expression. Proceedings of the National Academy of Sciences of the United States of America 100, 11255–11260.

Dwyer, M.A., Looger, L.L., Hellinga, H.W., 2004. Computational design of a biologically active enzyme. Science 304, 1967–1971.

Edinger, S., Cortis, C., Shenkin, P., Friesner, R., 1997. Solvation free energies of peptides: comparison of approximate continuum solvation models with accurate solution of the Poisson–Boltzmann equation. Journal of Physical Chemistry B 101, 1190–1197.

Emberly, E.G., Miller, J., Zeng, C., Wingreen, N.S., Tang, C., 2002a. Identifying proteins of high designability via surface-exposure patterns. Proteins: Structure, Function, and Bioinformatics 47, 295–304.

Emberly, E.G., Wingreen, N.S., Tang, C., 2002b. Designability of α-helical proteins. Proceedings of the National Academy of Sciences of the United States of America 99, 11163–11168.

Emberly, E.G., Mukhopadhyay, R., Tang, C., Wingreen, N.S., 2003. Flexibility of α-helices: results of a statistical analysis of database protein structures. Journal of Molecular Biology 327, 229–237.

Emberly, E.G., Mukhopadhyay, R., Tang, C., Wingreen, N.S., 2004. Flexibility of β-sheets: principal component analysis of database protein structures. Proteins: Structure, Function, and Bioinformatics 55, 91–98.

Eyrich, V.A., Standley, D.M., Felts, A.K., Friesner, R.A., 1999a. Protein tertiary structure prediction using a branch and bound algorithm. Proteins: Structure, Function, and Bioinformatics 35, 41–57.

Eyrich, V.A., Standley, D.M., Friesner, R.A., 1999b. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. Journal of Molecular Biology 288, 725–742.

Farinas, E., Regan, L., 1998. The de novo design of a rubredoxin-like Fe site. Protein Science 7, 1939–1946.

Finkelstein, A.V., Badretdinov, A.Y., Gutin, A.M., 1995. Why do proteins architectures have Boltzmann-like statistics? Proteins: Structure, Function, and Bioinformatics 23, 142–150.

Fiser, A., Do, R.K.G., Sali, A., 2000. Modeling of loops in protein structures. Protein Science 9, 1753–1773.

Fiser, A., Feig, M., Brooks-III, C.L., Sali, A., 2002. Evolution and physics in comparative protein structure modeling. Accounts of Chemical Research 35, 413–421.

Flöckner, H., Domingues, F.S., Sippl, M.J., 1997. Protein folds from pair interactions: a blind test in fold recognition. Proteins: Structure, Function, and Bioinformatics 1, 129–133.

Floudas, C.A., 1995. Nonlinear and Mixed-integer Optimization: Fundamentals and Applications. Oxford University Press, Oxford.

Floudas, C.A., 2000. Deterministic Global Optimization: Theory, Methods and Applications, Nonconvex Optimization and its Applications. Kluwer Academic Publishers, Dordrecht.

Forrest, L.R., Woolf, T.B., 2003. Discrimination of native loop conformations in membrane proteins: decoy library design and evaluation of effective energy scoring functions. Proteins: Structure, Function, and Bioinformatics 52, 491–509.

Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, New York.

Ghirlanda, G., Lear, J.D., Lombardi, A., DeGrado, W.F., 1998. From synthetic coiled coils to functional proteins: automated design of a receptor for the calmodulin-binding domain of calcineurin. Journal of Molecular Biology 281, 379–391.

Gillespie, B., Vu, D.M., Shah, P.S., Marshall, S.A., Dyer, R.B., Mayo, S.L., Plaxco, K.W., 2003. NMR and temperature-jump measurements of de novo designed proteins demonstrate rapid folding in the absence of explicit selection for kinetics. Journal of Molecular Biology 330, 813–819.

Godzik, A., Kolinski, A., Skolnick, J., 1993. De novo and inverse folding predictions of protein structure and dynamics. Journal of Computer-Aided Molecular Design 7, 397–438.

Goldstein, R.F., 1994. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophysical Journal 66, 1335–1340.

Gordon, D.B., Marshall, S.A., Mayo, S.L., 1999. Energy functions for protein design. Current Opinion in Structural Biology 9, 509–513.

Gordon, D.B., Hom, G.K., Mayo, S.L., Pierce, N.A., 2003. Exact rotamer optimization for protein design. Journal of Computational Chemistry 24, 232–243.

Güntert, P., Mumenthaler, C., Wüthrich, K., 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of Molecular Biology 273, 283–298.

Hao, M., Scheraga, H.A., 1999. Designing potential energy functions for protein folding. Current Opinion in Structural Biology 9, 184–188.

Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., Kim, P.S., 1998. High-resolution protein design with backbone freedom. Science 282, 1462–1467.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a K-means clustering algorithm. Applied Statistics 28, 100–108.

Hecht, M.H., Das, A., Go, A., Bradley, L.H., Wei, Y., 2004. De novo proteins from designed combinatorial libraries. Protein Science 13, 1711–1723.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M.J., 1990. Identification of native protein folds amongst a large number of incorrect models. Journal of Molecular Biology 216, 167–180.

Honig, B., Yang, A.S., 1995. Free energy balance in protein folding. Advances in Protein Chemistry 46, 27–58.

Honig, B., Sharp, K., Yang, A., 1993. Macroscopic models of aqueous solutions—biological and chemical applications. Journal of Physical Chemistry 97, 1101–1109.

Jacobson, M.P., Pincus, D.L., Rappa, C.S., Day, T.J.F., Honig, B., Shaw, D.E., Friesner, R.A., 2004. A hierarchical approach to all-atom protein loop prediction. Proteins: Structure, Function, and Bioinformatics 55, 351–367.

Jernigan, R.L., Bahar, I., 1996. Structure–derived potentials and protein simulations. Current Opinion in Structural Biology 6, 195–209.

Jones, D.T., 1997. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins: Structure Function and Bioinformatics S1, 185–191.

Jones, D.T., 1999a. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology 287, 797–815.

Jones, D.T., 1999b. Protein secondary structure prediction based on position specific scoring matrices. Journal of Molecular Biology 292, 195–202.

Jones, D.T., 2001. Predicting novel protein folds by using FRAGFOLD. Proteins: Structure Function and Bioinformatics S5, 127–132.

Jones, D.T., Guffin, L.J., 2003. Assembling novel protein folds from super-secondary structural fragments. Proteins: Structure, Function, and Bioinformatics 53, 480–485.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. A new approach to protein fold recognition. Nature 358, 86–89.

Karplus, P.A., 1997. Hydrophobicity regained. Protein Science 6, 1302–1307.

Karplus, K., Barret, C., Hughey, R., 1998. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14 (10), 846–856.

Karplus, K., Barret, C., Cline, M., Diekhans, M., Grate, L., Hughey, R., 1999. Predicting protein structure using only sequence information. Proteins: Structure, Function, and Bioinformatics S3, 121–125.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., Hughey, R., 2003. Combining local–structure, fold–recognition, and new fold methods for protein structure prediction. Proteins: Structure, Function, and Bioinformatics 53, 491–496.

Kelley, L.A., MacCallum, R.M., Sternberg, M.J.E., 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. Journal of Molecular Biology 299 (2), 499–520.

Kihara, D., Skolnick, J., 2003. The PDB is a covering set of small protein structures. Journal of Molecular Biology 334, 793–802.

Kim, D., Xu, D., Guo, J., Ellrott, K., Xu, Y., 2003. PROSPECT II: protein structure prediction program for genome-scale applications. Protein Engineering 16 (9), 641–650.

Klepeis, J.L., Floudas, C.A., 2002. Ab initio prediction of helical segments in polypeptides. Journal of Computational Chemistry 23, 245–266.

Klepeis, J.L., Floudas, C.A., 2003a. Prediction of beta-sheet topology and disulfide bridges in polypeptides. Journal of Computational Chemistry 24, 191–208.

Klepeis, J.L., Floudas, C.A., 2003b. Ab initio tertiary structure prediction of proteins. Journal of Global Optimization 25, 113–140.

Klepeis, J.L., Floudas, C.A., 2003c. ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. Biophysical Journal 85, 2119–2146.

Klepeis, J.L., Floudas, C.A., 2005. Analysis and prediction of loop segments in protein structure. Computers and Chemical Engineering 29, 423–436.

Klepeis, J.L., Pieja, M.T., Floudas, C.A., 2003a. A new class of hybrid global optimization algorithms for peptide structure prediction. Integrated hybrids. Computer Physics Communication 151, 121–140.

Klepeis, J.L., Pieja, M.T., Floudas, C.A., 2003b. Hybrid global optimization algorithms for protein structure prediction: alternating hybrids. Biophysical Journal 84, 869–882.

Klepeis, J.L., Floudas, C.A., Morikis, D., Tsokos, C.G., Argyropoulos, E., Spruce, L., Lambris, J.D., 2003c. Integrated computational and experimental approach for lead optimization and design of compstatin variants with improved activity. Journal of the American Chemical Society 125, 8422–8423.

Klepeis, J.L., Floudas, C.A., Morikis, D., Tsokos, C.G., Lambris, J.D., 2004. Design of peptide analogs with improved activity using a novel de novo protein design approach. Industrial and Engineering Chemistry Research 43, 3817–3826.

Klepeis, J.L., Wei, Y.N., Hecht, M.H., Floudas, C.A., 2005. Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double blind case study. Proteins: Structure, Function, and Bioinformatics 58, 560–570.

Koehl, P., Levitt, M., 1999a. De novo protein design. I. In search of stability and specificity. Journal of Molecular Biology 293, 1161–1181.

Koehl, P., Levitt, M., 1999b. De novo protein design. II. Plasticity in sequence space. Journal of Molecular Biology 293, 1183–1193.

Kono, H., Saven, J.G., 2001. Statistical theory for protein combinatorial libraries, packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. Journal of Molecular Biology 306, 607–628.

Kopp, J., Schwede, T., 2004. Automated protein structure homology modeling: a progress report. Pharmacogenomics Journal 5 (4), 405–416.

Kortemme, T., Morozov, A.V., Baker, D., 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. Journal of Molecular Biology 326, 1239–1259.

Kraemer-Pecore, C.M., Wollacott, A.M., Desjarlais, J.R., 2001. Computational protein design. Current Opinion in Chemical Biology 5, 690–695.

Kraemer-Pecore, C.M., Lecomte, J.T., Desjarlais, J.R., 2003. A de novo redesign of the WW domain. Protein Science 12, 2194–2205.

Kuhlman, B., Baker, D., 2000. Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences of the United States of America 97, 10383–10388.

Kuhlman, B., Baker, D., 2004. Exploring folding free energy landscapes using computational protein design. Current Opinion in Structural Biology 14, 89–95.

Kuhlman, B., O'Neill, J.W., Kim, D.E., Zhang, K.Y.J., Baker, D., 2002. Accurate computer-based design of a new backbone conformation in the second turn of protein l. Journal of Molecular Biology 315, 471–477.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., Baker, D., 2003. Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–1368.

Larson, S.M., England, J.L., Desjarlais, J.R., Pande, V.S., 2002. Thoroughly sampling sequence space: large-scale protein design of structural ensembles. Protein Science 11, 2804–2813.

Lathrop, R.H., 1994. The protein threading problem with sequence amino–acid interaction preferences is NP-complete. Protein Engineering 7 (9), 1059–1068.

Lee, C., 1994. Predicting protein mutant energetics by self-consistent ensemble optimization. Journal of Molecular Biology 236, 918–939.

Lee, J., Scheraga, H.A., 1999. Conformational space annealing by parallel computations: extensive conformational search of met-enkephalin and the 20-residue membrane-bound portion of melittin. International Journal of Quantum Chemistry 75, 255–265.

Lee, J., Scheraga, H.A., Rackovsky, S., 1997. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. Journal of Computational Chemistry 18, 1222–1232.

Lee, J., Scheraga, H.A., Rackovsky, S., 1998. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. Biopolymers 46, 103–115.

Lee, J., Pillardy, J., Czaplewski, C., Arnautova, Y., Ripoll, D.R., Liwo, A., Gibson, K.D., Wawak, R.J., Scheraga, H.A., 2000. Efficient parallel algorithms in global optimization of potential energy functions for peptides, proteins and crystals. Computer Physics Communication 128, 399–411.

Lee, J., Ripoll, D.R., Czaplewski, C., Pillardy, J., Wedemeyer, W.J., Scheraga, H.A., 2001. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. Journal of Physical Chemistry B 105, 7291–7298.

Lee, J., Kim, S.-Y., Joo, K., Kim, I., Lee, J., 2004. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins: Structure Function and Bioinformatics 56, 704–714.

Lesser, G.J., 1990. Hydrophobicity of amino acid subgroups in proteins. Proteins: Structure Function and Bioinformatics 8, 6–13.

Levinthal, C., 1969. How to fold graciously. In: Debrunner, P., Tsibris, J.C.M., Münck, E.M. (Eds.), Mossbauer Spectroscopy in Biological Systems. University of Illinois Press, Urbana, pp. 22–24.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S., Tsai, J., 1997. Protein folding: the endgame. Annual Review of Biochemistry 66, 549–579.

Li, H., Tang, C., Wingreen, N.S., 2002. Designability of protein structures: a lattice-model study using the Miyazawa–Jernigan matrix. Proteins: Structure, Function, and Bioinformatics 49, 403–412.

Li, X., Jacobson, M.P., Friesner, R.A., 2004. High-resolution prediction of protein helix positions and orientations. Proteins: Structure Function and Bioinformatics 55, 368–382.

Lim, W.A., Hodel, A., Sauer, R.T., Richards, F.M., 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. Proceedings of the National Academy of Sciences of the United States of America 91, 423–427.

Liu, J., Rost, B., 2002. Target space for structural genomics revisited. Bioinformatics 18 (7), 922–933.

Liu, J., Hegyi, H., Acton, T.B., Montelione, G.T., Rost, B., 2004. Automatic target selection for structural genomics on eukaryotes. Proteins: Structure, Function, and Bioinformatics 56, 188–200.

Liwo, A., Oldziej, S., Pincus, M.R., Wawak, R.J., Rackovsky, S., Scheraga, H.A., 1997a. A united-residue force field for off-lattice protein structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. Journal of Computational Chemistry 18, 849–873.

Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S., Oldziej, S., Scheraga, H.A., 1997b. A united-residue force field for off-lattice protein structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by z-score optimization. Journal of Computational Chemistry 18, 874–887.

Liwo, A., Odziej, S., Kamierkiewicz, R., Groth, M., Czaplewski, C., 1997c. Design of a knowledge-based force field for off-lattice simulations of protein structure. Acta Biochimica Polonica 44, 527–547.

Liwo, A., Kamierkiewicz, R., Czaplewski, C., Groth, M., Odziej, S., Wawak, R.J., Rackovsky, S., Pincus, M.R., Scheraga, H.A., 1998. United-residue force field for off-lattice protein-structure simulations: III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. Journal of Computational Chemistry 19, 259–276.

Liwo, A., Czaplewski, C., Pillardy, J., Scheraga, H.A., 2001. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. Journal of Chemical Physics 115, 2323–2347.

Liwo, A., Arlukowicz, P., Czaplewski, C., Oldziej, S., Pillardy, J., Scheraga, H.A., 2002. A method for optimizing potential-energy functions by hierarchical design of the potential-energy landscape: application to the UNRES force field. Proceedings of the National Academy of Sciences of the United States of America 99, 1937–1942.

Liwo, A., Odziej, S., Czaplewski, C., Kozlowska, U., Scheraga, H.A., 2004. Parametrization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. Journal of Physical Chemistry B 108, 9421–9438.

Looger, L.L., Hellinga, H.W., 2001. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. Journal of Molecular Biology 307, 429–445.

Looger, L.L., Dwyer, M.A., Smith, J.J., Hellinga, H.W., 2003. Computational design of receptor and sensor proteins with novel functions. Nature 423, 185–190.

Loose, C., Klepeis, J.L., Floudas, C.A., 2004. A new pairwise folding potential based on improved decoy generation and side-chain packing. Proteins: Structure, Function, and Bioinformatics 54, 303–314.

Lovejoy, B., Choe, S., Cascio, D., McRorie, D.K., DeGrado, W.F., Eisenberg, D., 1993. Crystal structure of a synthetic triple-stranded α-helical bundle. Science 259, 1288–1293.

Lu, H., Skolnick, J., 2001. A distance-dependent knowledge-based potential for improved protein structure selection. Proteins: Structure, Function, and Bioinformatics 44, 223–232.

MacKerell Jr., A.D., Bashford, D., Bellott, M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher III, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., Karplus, M., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B 102, 3586–3616.

Maiorov, V.N., Crippen, G.M., 1992. Contact potential that recognizes the correct folding of globular proteins. Journal of Molecular Biology 227, 876–888.

Malakauskas, S.M., Mayo, S.L., 1998. Design, structure, and stability of a hyperthermophilic protein variant. Nature Structural Biology 5, 470–475.

Mallik, B., Katragadda, M., Spruce, L.A., Carafides, C., Tsokos, C.G., Morikis, D., Lambris, J.D., 2005. Design and NMR characterization of active analogues of compastin containing non-natural amino acids. Journal of Medicinal Chemistry 48, 274–286.

Meller, J., Wagner, M., Elber, R., 2002. Maximum feasibility guideline in the design and analysis of protein folding potentials. Journal of Computational Chemistry 23, 111–118.

Miller, J., Zeng, C., Wingreen, N.S., Tang, C., 2002. Emergence of highly designable protein-backbone conformations in an off-lattice model. Proteins: Structure, Function, and Bioinformatics 47, 506–512.

Miyazawa, S., Jernigan, R.L., 1999. An emperical energy potential with a reference state for protein fold and sequence recognition. Proteins: Structure, Function, and Bioinformatics 36, 357–369.

Momany, F.A., McGuire, R.F., Burgess, A.W., Scheraga, H.A., 1975. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. Journal of Physical Chemistry 79, 2361–2381.

Mönnigmann, M., Floudas, C.A., 2005. Protein loop structure prediction with flexible stem geometries, submitted for publication.

Mooers, B.H.M., Datta, D., Baase, W.A., Zollars, E.S., Mayo, S.L., Matthews, B.W., 2003. Repacking the core of T4 lysozyme by automated design. Journal of Molecular Biology 332, 741–756.

Moult, J., 1999. Predicting protein three-dimensional structure. Current Opinion in Biotechnology 10, 583–588.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., Pedersen, J.T., 1997. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins: Structure, Function, and Bioinformatics S1, 2–6.

Moult, J., Fidelis, K., Zemla, A., Hubbard, T., 2001. Critical assessment of methods of protein structure prediction (CASP)—Round 4. Proteins: Structure, Function, and Bioinformatics S5, 2–7.

Moult, J., Fidelis, K., Zemla, A., Hubbard, T., 2003. Critical assessment of methods of protein structure prediction (CASP)—Round V. Proteins: Structure, Function, and Bioinformatics 53, 334–339.

Munoz, V., Thompson, P.A., Hofrichter, J., Eaton, W.A., 1997. Folding dynamics and mechanism of beta-hairpin formation. Nature 390, 196–199.

Murzin, A.G., 2004. Protein structure watch: making "predictions" easy. www.forcasp.org.

Murzin, A.G., Lesk, A.M., Chothia, C., 1994a. Principles determining the structure of beta-sheet barrels in proteins. i. a theoretical analysis. Journal of Molecular Biology 236, 1369–1381.

Murzin, A.G., Lesk, A.M., Chothia, C., 1994b. Principles determining the structure of beta-sheet barrels in proteins. ii. the observed structures. Journal of Molecular Biology 236, 1382–1400.

Némethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S., Scheraga, H.A., 1992. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm with application to proline-containing peptides. Journal of Physical Chemistry 96, 6472–6484.

Notredame, C., 2002. Recent progress in multiple sequence alignment: a survey. Pharmacogenomics Journal 3 (1), 131–144.

Novotny, J., Rashin, A.A., Bruccoleri, R.E., 1984. Criteria that discriminate between native proteins and incorrectly folded models. Proteins: Structure, Function, and Bioinformatics 177, 788–818.

Ohkubo, Y.Z., Crippen, G.M., 2000. Potential energy functions for continuous state models of globular proteins. Journal of Computational Biology 7, 363–379.

Orengo, C.A., Jones, D.T., Thornton, J.M., 1994. Protein superfamilies and domain superfolds. Nature 372, 631–634.

Pabo, C., 1983. Molecular technology. Designing proteins and peptides. Nature 301 (5897), 200.

Pande, V.S., Rokhsar, D.S., 1999. Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. Proceedings of the National Academy of Sciences of the United States of America 96, 9062–9067.

Park, S., Yang, X., Saven, J.G., 2004. Advances in computational protein design. Current Opinion in Structural Biology 14, 487–494.

Pierce, N.A., Winfree, E., 2002. Protein design is NP-hard. Protein Engineering 15, 779–782.

Pierce, N.A., Spriet, J.A., Desmet, J., Mayo, S.L., 2000. Conformational splitting: a more powerful criterion for dead-end elimination. Journal of Computational Chemistry 21, 999–1009.

Pillardy, J., Czaplewski, C., Liwo, A., Wedemeyer, W.J., Lee, J., Ripoll, D.R., Arlukowicz, P., Oldziej, S., Arnautova, E.A., Scheraga, H.A., 2001. Development of physics-based energy functions that predict medium resolution structure for proteins of α, β and α/β structural classes. Journal of Physical Chemistry B 105, 7299–7311.

Pokala, N., Handel, T.M., 2001. Review: protein design—where we were, where we are, where we're going. Journal of Structural Biology 134, 269–281.

Ponder, J.W., Richards, F.M., 1987a. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. Journal of Computational Chemistry 8, 1016–1024.

Ponder, J.W., Richards, F.M., 1987b. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of Molecular Biology 193, 775–791.

Przybylski, D., Rost, B., 2004. Improving fold recognition without folds. Journal of Molecular Biology 341, 255–269.

Radzicka, A., Wolfenden, R., 1988. Comparing the polarities of amino acids: side chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemical Journal 27, 1664–1670.

Richards, F.M., Hellinga, H.W., 1991. Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. Journal of Molecular Biology 222, 763–785.

Richards, F.M., Hellinga, H.W., 1994. Optimal sequence selection in proteins of known structure by simulated evolution. Proceedings of the National Academy of Sciences of the United States of America 91, 5803–5807.

Richards, F.M., Caradonna, J.P., Hellinga, H.W., 1991. Construction of new ligand binding sites in proteins of known structure. II. Grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin. Journal of Molecular Biology 222, 787–803.

Ripoll, D., Liwo, A., Scheraga, H.A., 1998. New developments of the electrostatically driven Monte Carlo method: tests on the membrane-bound portion of melittin. Biopolymers 46, 117–126.

Rohl, C.A., Strauss, C.E.M., Chivian, D., Baker, D., 2004. Modeling structurally variable regions in homologous proteins with Rosetta. Proteins: Structure, Function, and Bioinformatics 55, 656–677.

Ross, S.A., Sarisky, C.A., Su, A., Mayo, S.L., 2001. Designed protein G core variants fold to native-like structures: sequence selection by ORBIT tolerates variation in backbone specification. Protein Science 10, 450–454.

Rost, B., 2001. Review: protein secondary structure prediction continues to rise. Journal of Structural Biology 134, 204–218.

Samudrala, R., Moult, J., 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275, 895–916.

Saven, J.G., 2001. Designing protein energy landscapes. Chemical Reviews 101, 3113–3130.

Saven, J.G., 2002. Combinatorial protein design. Current Opinion in Structural Biology 2, 453–458.

Saven, J.G., 2003. Connecting statistical and optimized potentials in protein folding via a generalized foldability criterion. Journal of Chemical Physics 118, 6133–6136.

Sayle, R., Milner-White, E.J., 1995. RasMol: biomolecular graphics for all. Trends in Biochemical Science 20, 374.

Scott, W.R.P., Hunenberger, P.H., Trioni, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., Kruger, P., VanGunsteren, W.F., 1997. The GROMOS biomolecular simulation program package. Journal of Physical Chemistry A 103, 3596–3607.

Shakhnovich, E.I., 1998. Protein design: a perspective from simple tractable models. Folding and Design 3, 45–58.

Shao, Y., Bystroff, C., 2003. Predicting interresidue contacts using templates and pathways. Proteins: Structure, Function, and Bioinformatics 53, 497–502.

Shi, J., Tom, L.B., Kenji, M., 2001. FUGUE: sequence-structure homology prediction using environment-specific substitution tables and structure-dependent gap penalties. Journal of Molecular Biology 310, 243–257.

Shimaoka, M., Shifman, J.M., Jing, H., Takagi, L., Mayo, S.L., Springer, T.A., 2000. Computational design of an integrin I domain stabilized in the open high affinity conformation. Nature Structural Biology 7, 674–678.

Sia, S.K., Kim, P.S., 2001. A designed protein with packing between left-handed and right-handed helices. Biochemistry 40, 8981–8989.

Simons, K.T., Kooperberg, C., Huang, C., Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of Molecular Biology 268, 209–225.

Simons, K.T., Ruczinki, I., Kooperberg, C., Fox, B.A., Bystroff, C., Baker, D., 1999. Improved recognition of native-like structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins: Structure, Function, and Bioinformatics 34, 82–95.

Skolnick, J., Jaroszewski, L., Kolinski, A., Godzik, A., 1997a. Derivation and testing of pair potential for protein folding. When is quasichemical approximation correct? Protein Science 6, 676–688.

Skolnick, J., Kolinski, A., Oritz, A.R., 1997b. MONSSTER: a method for folding globular proteins with a small number of distance constraints. Journal of Molecular Biology 265, 217–241.

Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P., Boniecki, M., 2001. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering and structure refinement. Proteins: Structure, Function, and Bioinformatics 5 (Suppl.), 149–156.

Skolnick, J., Zhang, Y., Arakaki, A.K., Kolinski, A., Boniecki, M., Szilágyi, A., Kihara, D., 2003. TOUCHSTONE: a unified approach to protein structure prediction. Proteins: Structure, Function, and Bioinformatics 53, 469–479.

Skolnick, J., Kihara, D., Zhang, Y., 2004. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins: Structure, Function, and Bioinformatics 56, 502–518.

Srinivasan, R., Rose, G.D., 1995. LINUS: a hierarchic procedure to predict the fold of a protein. Proteins: Structure, Function, and Bioinformatics 22, 81–89.

Srinivasan, R., Rose, G.D., 2002. Ab initio prediction of protein structure using LINUS. Proteins: Structure, Function, and Bioinformatics 47, 489–495.

Street, A.G., Mayo, S.L., 1999. Computational protein design. Structure With Folding & Design 7, R105–R109.

Strop, P., Mayo, S.L., 1999. Rubredoxin variant folds without irons. Journal of the American Chemical Society 121, 2341–2345.

Su, A., Mayo, S.L., 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Science 6, 1701–1707.

Tanaka, S., Scheraga, H.A., 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9, 945–950.

Tang, Y., Ghirlanda, G., Vaidehi, N., Kua, J., Mainz, D.T., Goddard, I.W., DeGrado, W.F., Tirrell, D.A., 2001. Stabilization of coiled-coil peptide domains by introduction of trifluoroleucine. Biochemistry 40, 2790–2796.

Tobi, D., Elber, R., 2000. Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins: Structure, Function, and Bioinformatics 41, 40–46.

Tobi, D., Shafran, G., Linial, N., Elber, R., 2000. On the design and analysis of protein folding potentials. Proteins: Structure, Function, and Bioinformatics 40, 71–85.

Tramontano, A., Morea, V., 2003. Assessment of homology-based predictions in CASP5. Proteins: Structure, Function, and Bioinformatics 53, 352–368.

Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A., Baker, D., 2003. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins: Structure, Function, and Bioinformatics 52, 76–87.

Vendruscolo, M., Domany, E., 1998. Pairwise contact potentials are unsuitable for protein folding. Journal of Chemical Physics 109, 11101–11108.

Vendruscolo, M., Najmanovich, R., Domany, E., 2000. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins: Structure, Function, and Bioinformatics 38, 134–148.

Vitkup, D., Melomud, E., Moult, J., Sander, C., 2001. Completeness in structural genomics. Nature Structural Biology 8 (6), 559–566.

Voigt, C.A., Mayo, S.L., Arnold, F.H., Wang, Z.G., 2001. Computational method to reduce the search space for directed protein evolution. Proceedings of the National Academy of Sciences of the United States of America 98 (7), 3778–3783.

Walsh, S.T., Cheng, H., Bryson, J.W., Roder, H., DeGrado, W.F., 1999. Solution structure and dynamics of a de novo designed three–helix bundle protein. Proceedings of the National Academy of Sciences of the United States of America 96, 5486–5491.

Wang, Y., Zhang, H., Li, W., Scott, R.A., 1995. Discriminating compact non-native structures from the native structures of globular proteins. Proceedings of the National Academy of Sciences of the United States of America 92, 709–713.

Watters, A.L., Baker, D., 2004. Searching for folded proteins in vitro and in silico. European Journal of Biochemistry 271, 1615–1622.

Wernisch, L., Hery, S., Wodak, S.J., 2000. Automatic protein design with all atom force-fields by exact and heuristic optimization. Journal of Molecular Biology 301, 713–736.

Wilson, C., Mace, J.E., Agard, D.A., 1991. Computational method for the design of enzymes with altered substrate specificity. Journal of Molecular Biology 220, 495–506.

Wisz, M.S., Hellinga, H.W., 2003. An empirical model for electrostatic interactions in proteins incorporating multiple geometry–dependent dielectric constants. Proteins: Structure, Function, and Bioinformatics 51, 360–377.

Xia, Y., Huang, E.S., Levitt, M., Samudrala, R., 2000. Ab initio construction of protein tertiary structure using a hierarchical approach. Journal of Molecular Biology 300, 171–185.

Xiang, Z., Soto, C., Honig, B., 2002. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proceedings of the National Academy of Sciences of the United States of America 99, 7432–7437.

Xu, J., Li, M., 2003. Assessment of RAPTOR's linear programming approach in CAFASP3. Proteins: Structure, Function, and Bioinformatics 53, 579–584.

Xu, J., Li, M., Kim, D., Xu, Y., 2003. RAPTOR: optimal protein threading by linear programming. Journal of Bioinformatics and Computational Biology 1, 95–117.

Xu, Y., Xu, D., 2000. Protein threading using PROSPECT: design and evolution. Proteins: Structure, Function, and Bioinformatics 40, 343–354.

Zhang, C., Liu, S., Zhou, Y., 2003. Accurate and efficient loop selections by the DFIRE—based all atom statistical potential. Protein Science 13, 391–399.

Zhang, C., Liu, S., Zhou, H., Zhou, Y., 2004. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Science 13, 400–411.

Zhang, L., Skolnick, J., 1998. How do potentials derived from structural database relate to "true" potentials? Protein Science 7, 112–122.

Zhang, Y., Skolnick, J., 2004a. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophysical Journal 87, 2647–2655.

Zhang, Y., Skolnick, J., 2004b. Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America 101 (20), 7594–7599.

Zhang, Y., Skolnick, J., 2004c. SPICKER: a clustering approach to identify near-native protein folds. Journal of Computational Chemistry 25, 865–871.

Zhang, Y., Kolinski, A., Skolnick, J., 2003. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophysical Journal 85, 1145–1164.

Zhou, H., Zhou, Y., 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 11, 2714–2726.

Zhu, Y., Alonso, D.O., Maki, K., Huang, C.Y., Lahr, S.J., Daggett, V., Roder, H., DeGrado, W.F., Gai, F., 2003. Ultrafast folding of alpha3D: a de novo designed three-helix bundle protein. Proceedings of the National Academy of Sciences of the United States of America 100, 15486–15491.

Zou, J., Saven, J.G., 2000. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. Journal of Molecular Biology 296, 281–294.

Zwanzig, R., Szabo, A., Bagchi, B., 1992. Levinthal's paradox. Proceedings of the National Academy of Sciences of the United States of America 89, 20–22.